J. R. Statist. Soc. A (1994)
157, *Part* 3, *pp.* 357–416

Bayesian Approaches to Randomized Trials

By DAVID J. SPIEGELHALTER†, LAURENCE S. FREEDMAN

Medical Research Council Biostatistics Unit, Cambridge, UK

National Cancer Institute, Bethesda, USA

and MAHESH K. B. PARMAR

Medical Research Council Cancer Trials Office, Cambridge, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Medical Section *on Wednesday, February 16th, 1994, the President*, Professor D. J. Bartholomew, *in the Chair*]

SUMMARY

Statistical issues in conducting randomized trials include the choice of a sample size, whether to stop a trial early and the appropriate analysis and interpretation of the trial results. At each of these stages, evidence external to the trial is useful, but generally such evidence is introduced in an unstructured and informal manner. We argue that a Bayesian approach allows a formal basis for using external evidence and in addition provides a rational way for dealing with issues such as the ethics of randomization, trials to show treatment equivalence, the monitoring of accumulating data and the prediction of the consequences of continuing a study. The motivation for using this methodology is practical rather than ideological.

Keywords: CLINICAL TRIALS; ETHICS; POWER; PREDICTION; PRIOR DISTRIBUTION; RANGE OF EQUIVALENCE; SHRINKAGE; STOPPING RULES; SUBJECTIVE PROBABILITIES

1. INTRODUCTION

The accepted statistical techniques for the design, monitoring and reporting of controlled clinical trials are based on the frequentist theory of hypothesis testing developed by Neyman and Pearson (see, for example, Pocock (1983)). The impact of the frequentist approach is reflected in the established use of type I and type II error for clinical trial design and p-values and confidence intervals for analysis, and in the demand by both medical journals and regulatory authorities for reporting within this framework.

Our paper is about the use of an alternative, Bayesian, theory of statistical inference in clinical trials. Clinical investigation is an essentially dynamic process, in which any individual study takes place in a context of continuously increasing knowledge. New information emerges not only from a trial as it progresses but also from other studies that are relevant to the questions addressed by the trial. Furthermore, it is unlikely that a single trial will provide a definitive clinical conclusion, whether in industry in which drug registration is the goal, or in publicly funded research in which influencing clinical practice is the main objective. It therefore seems vital to acknowledge the *context* in which a study takes place, and to emphasize that the data from a single trial report *add* to available evidence, rather than form the basis for decision-making

† *Address for correspondence*: Medical Research Council Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, CB2 2SR, UK.

 0035–9238/94/157357

in themselves. The Bayesian method is naturally dynamic, in that prior distributions of belief regarding treatment differences may be modified by new data in a formal way by using Bayes theorem.

Such an approach is standard in diagnostic testing (Ingelfinger *et al.*, 1987). The practical value of a diagnostic test cannot be decided on the basis of its sensitivity and specificity alone—to make decisions we need the prior probability of disease, and the perceived threshold of belief that would make intervention worthwhile. We follow others (e.g. Peto *et al.* (1976)) in arguing that the same is true for clinical trials: the conclusions to be drawn from a classically positive or negative result depend on the degree of scepticism concerning the efficacy of the new treatment based on evidence external to the study, and the level of improvement required before it is considered clinically superior.

Bayesian methods will not, however, be adopted in the absence of a clear demonstration of the practical advantages which accrue from their use. Moreover their practical application requires that the statistician moves into statistical territories which may be unfamiliar, particularly the appropriate choice of a prior distribution. The aim of our paper is therefore twofold: first, to highlight some advantages of Bayesian methods over current practice in clinical trials and, second, to familiarize statisticians with the concepts and practical methods involved in their application. In Section 2 we comment briefly on the classical hypothesis testing approach to clinical trials, and in Section 3 we give an overview of the alternative Bayesian framework using the simplest formulation. In Sections 4 and 5 we discuss prior distributions, analysis and monitoring, including suggestions for formal specification of priors representing scepticism and enthusiasm. Section 6 deals with possible roles for prediction methods with sample size calculation as a special case, while Section 7 contains applications to real trials. In Section 8 we discuss issues arising in monitoring and reporting of results. Some technical details are given in Appendixes A and B.

The practical emphasis of this paper means that Bayesian techniques will not be claimed to be superior simply on the basis of foundational arguments. We do not consider alternative trial designs such as adaptive allocation rules, and we fully accept the need for randomization. However, we do question the standard basis for drawing inferences following a classical randomized trial, and the next section introduces some of our misgivings.

2. HYPOTHESIS TESTING AND CLINICAL TRIALS

2.1. *Null Hypothesis and Ranges of Equivalence*

Suppose that a trial is comparing two treatments, which we shall term the 'new' and the 'old', and that the true treatment difference is summarized by a parameter δ, where large values of δ correspond to superiority of the new treatment. Statistical analysis usually centres on a significance test of a null hypothesis H_0: $\delta = 0$. However, in some circumstances the treatments may be so unequal in their 'costs', e.g. in their toxicity, inconvenience or monetary cost, that it is commonly accepted that the more costly treatment will be required to achieve at least a certain margin of benefit, δ_I, before it can be even considered: hence $\delta < \delta_I$ corresponds to clinical inferiority of the new treatment (Schwartz *et al.*, 1980). Another value, δ_S, may be postulated, where $\delta > \delta_S$ indicates clinical superiority of the new treatment: δ_S is

sometimes termed the 'minimal clinically worthwhile benefit'. We call (δ_I, δ_S) the *range of equivalence* (Freedman *et al.*, 1984), such that if we were certain that δ lay within this interval we would be unable to make a definitive choice of treatment. Clearly the specification of δ_I and δ_S is not straightforward, but the concept is becoming recognized as being both feasible and useful (Armitage, 1989; Fleming and Watelet, 1989). We note that it is quite reasonable that the range of equivalence will change as a trial progresses.

As we shall discuss further in Section 5.1, conclusions may be drawn by relating evidence on δ to the three intervals $\delta < \delta_I$, $\delta_I < \delta < \delta_S$ and $\delta > \delta_S$. This formulation also includes trials that aim to demonstrate equivalence of treatments (Dunnett and Gent, 1977; Makuch and Simon, 1982), where we may set the range of equivalence as $(-\delta_E, +\delta_E)$ for a specified δ_E.

2.2. *Alternative Hypothesis*

Current statistical practice in clinical trial design is to formulate a point alternative hypothesis H_1, and a sample size is chosen which guarantees, under H_1, an acceptably high statistical power (e.g. 90%) of rejecting H_0 by using a specified statistical test at a given significance level (e.g. 0.05). Usually H_0 corresponds to $\delta = 0$. Medical statisticians continue to use different prescriptions for specifying the value δ_A which should represent H_1 (Spiegelhalter and Freedman, 1986). The main variants are that δ_A should represent

(a) the smallest clinically worthwhile difference (Lachin, 1981), or
(b) a difference that the investigator thinks is 'worth detecting' (Fleiss, 1981) or
(c) a difference that is thought likely to occur (Halperin *et al.*, 1982).

The first variant corresponds to taking δ_A as δ_S, whereas the second leaves the choice to the investigator, who may well choose δ_A to be unrealistically large to reduce the required sample size. Approach (c) differs in stressing the importance of *plausibility* as a concept for choosing H_1 (Peto *et al.*, 1976). Suppose that we set δ_A as the *expectation* of the benefit of a new treatment. In practice, clinicians participating in trials often have such expectations close to the *demands* that they would make of the treatment before using it routinely—in our notation, δ_A will be near δ_S. This similarity provides a sound basis for randomization but often leads to confusion between expectations and demands.

In heart disease and cancer research, trials with much larger sample sizes than were previously common have now been conducted to detect relatively small but clinically worthwhile treatment differences, using the argument that only such relatively small differences are likely (Yusuf *et al.*, 1984). Having accepted the role of the plausibility of a given improvement, it is a short step to an explicit expression of prior belief, and to a Bayesian perspective.

2.3. *Sequential Analysis*

Interim analyses are becoming increasingly popular both in public sector and pharmaceutical industry trials, with an accompanying large body of statistical literature: see, for example, Jennison and Turnbull (1990) and Whitehead (1992). Within the frequentist framework it is necessary to specify a stopping rule before the trial analysis

begins, spending the overall type I error at the interim analyses through the course of the trial. The formal use of these methods raises many serious conceptual and practical difficulties (Cornfield, 1966; Berry, 1987; Freedman and Spiegelhalter, 1989), even if they are only considered to be stopping guidelines rather than strict rules.

3. BAYESIAN FRAMEWORK

The basic paradigm of Bayesian statistics is straightforward. Initial beliefs concerning a parameter of interest, which could be based on objective evidence or subjective judgment or a combination, are expressed as a prior distribution. Evidence from further data is summarized by a likelihood function for the parameter, and the normalized product of the prior and the likelihood form the posterior distribution on the basis of which conclusions should be drawn (see, for example, Lee (1987)). In this paper we shall illustrate each of these steps with practical examples, and later in Section 6 we emphasize additional properties related to prediction of future outcomes.

This inferential process is often extended to a theory for decision-making by the introduction of utilities for certain outcomes. Although there have been many attempts to place clinical trials within such a decision theoretic framework, in our formulation we specifically do not include utility assessments. Our reason (Spiegelhalter and Freedman, 1988) is that when the decision is whether or not to discontinue the trial, coupled with whether or not to recommend one treatment in preference to the other, the consequences of any particular course of action are so uncertain that they make the meaningful specification of utilities rather speculative. Since we are dealing with a society with considerable freedom of choice of treatment, the implications of reporting a 'conclusive' trial result cannot currently be accurately modelled.

3.1. *Prior to Posterior Analysis*

To emphasize conceptual rather than technical issues we deliberately present only the simplest analysis which can be carried out without specialist software. More elaborate analyses are briefly discussed, although our formulation can cope with a wide variety of problems. Throughout this paper we use the notation that $p(\ )$ and $\phi(\ )$ are density functions.

3.1.1. *Likelihood*

We shall assume that our data after m observations can be summarized by a statistic x_m, whose distribution is

$$p(x_m) = \phi(x_m \mid \delta, \sigma^2/m), \tag{1}$$

where ϕ represents a Gaussian distribution with mean δ and variance σ^2/m, δ is the parameter of interest and σ^2 is assumed known: in comparative trials δ is the true treatment difference and x_m is the sample difference.

This assumption of a normal likelihood covers many situations: if individual responses are assumed Gaussian with variance $\sigma^2/2$, δ is the true difference in mean response, and x_m is the difference in group sample means where m individuals are allocated to each treatment; in survival analysis with proportional hazards, if m is the total number of events observed, $x_m = 4L_m/m$ where L_m is the log-rank test

statistic (observed − expected events in a treatment group), and δ is the log-hazard ratio, then x_m has approximately a distribution given by equation (1) with $\sigma^2 = 4$ (Tsiatis, 1981). For rare events, we have a similar approximation in which δ is the log-odds ratio, m is the number of events, x_m is the observed log-odds ratio and $\sigma^2 = 4$. For binomial responses with higher event rates, x_m is the difference in sample response rates and, strictly speaking, σ^2 depends on the unknown response rates, but in this case and in that of normal responses with unknown variance an estimate of σ^2 may be used in equation (1) which for sufficiently large m will be adequate. Rate ratios can also be handled within this framework (Pocock and Hughes, 1989). Whitehead (1992) exploited the use of normal likelihoods for efficient score statistics as the basis for sequential monitoring, and showed the wide applicability of this approach.

3.1.2. *Prior distribution*

We denote the prior distribution $p_0(\delta)$, indicating our belief having made zero observations in our trial. With a normal likelihood it is mathematically convenient, and often reasonably realistic, to make the assumption that $p_0(\delta)$ has the form

$$p_0(\delta) = \phi(\delta \mid \delta_0, \sigma^2/n_0) \tag{2}$$

where δ_0 is the prior mean. This prior is equivalent to a normalized likelihood arising from a (hypothetical) trial of n_0 patients with an observed value δ_0 of the treatment difference statistic. We shall use this normal assumption for the expressions shown below and in our examples: although it could be argued that a symmetric distribution will not adequately reflect opinion about a treatment benefit, it should be remembered that it is only the shape of the prior in the area supported by the likelihood that is important.

In general, to accommodate asymmetric distributions, we have found it convenient to use 'mixture' priors with two normal components, each of which may have an upper or lower bound. Either of the components may have zero variance, formally corresponding to $n_0 = \infty$, thus making a lump of probability on δ_0. The resulting family is both flexible and mathematically tractable (see Appendix A). We note that Carlin *et al.* (1993) describe more elaborate analyses within a proportional hazards regression model.

3.1.3. *Posterior distribution*

For the prior given in equation (2) and likelihood (1), we obtain by Bayes theorem a posterior distribution

$$\begin{aligned} p_m(\delta) &\propto p(x_m \mid \delta)\, p_0(\delta) \\ &= \phi\left(\delta \,\middle|\, \frac{n_0\delta_0 + mx_m}{n_0 + m}, \frac{\sigma^2}{n_0 + m}\right) \end{aligned} \tag{3}$$

where the subscript indicates our belief after m observations. The posterior mean serves as a point estimate, while $100\gamma\%$ credible interval estimates take the form of regions I_γ such that $\int_{I_\gamma} p_m(\delta)\, d\delta = \gamma$. For equation (3), 95% estimates, for example, are formed simply from the posterior mean ± 1.96 posterior standard deviations.

3.2. *Example 1: Medical Research Council Neutron Therapy Trial*

3.2.1. *The trial*

Errington *et al.* (1991) reported a trial of high energy neutron therapy for treatment of pelvic cancers. An *ad hoc* independent data monitoring committee was set up by the Medical Research Council (MRC) in January 1990 to review the interim results after 151 patients with locally advanced cancers had received either neutron therapy (90 patients) or conventional radiotherapy (61 patients).

3.2.2. *Range of equivalence*

Interviews were conducted in March 1988 with 10 selected clinicians and physicists, knowledgeable about neutron therapy, before the disclosure of any of the trial results. On average, respondents reported that they would require a change in one-year survival from 50% (assumed for standard photon therapy) to 61.5% for neutron therapy to be recommended as routine treatment. Under a proportional hazards assumption such a change corresponds to a hazard ratio of $\log 0.50/\log 0.615 = 1.426$ against photon therapy. Using the log-hazard scale, we therefore take $\log 1.426 = 0.355$ as our upper limit of the range of equivalence, δ_S, and $\delta_I = 0$ as our lower limit: each is shown as a dotted line in Fig. 1.

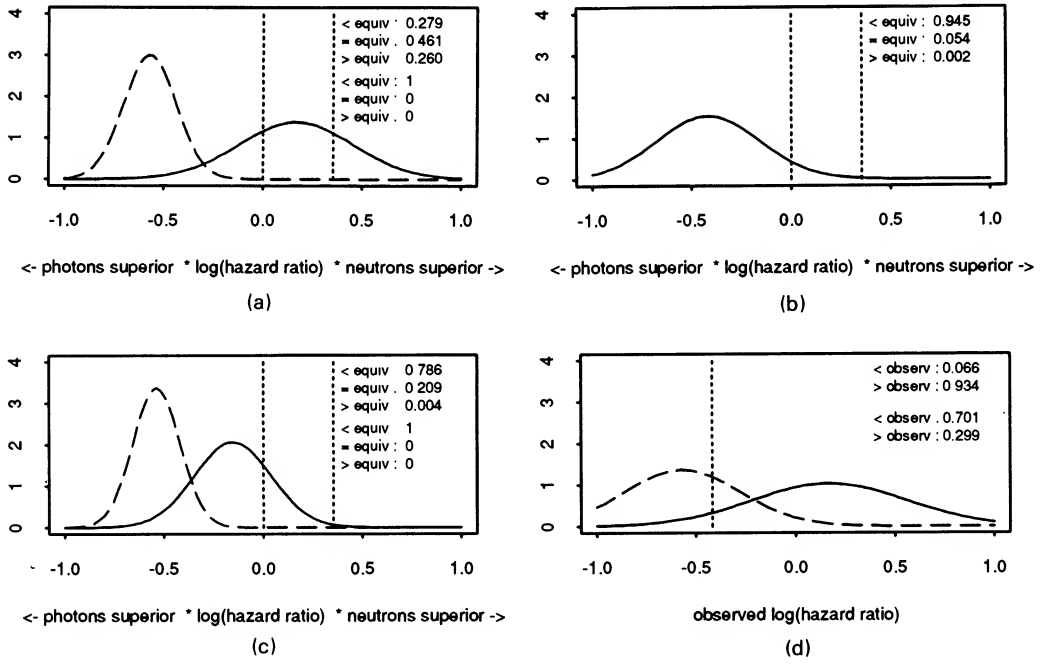

Fig. 1. Neutron therapy prior, likelihood and posterior distributions (the predictive distributions are discussed in Section 6.3) (———, prior based on clinical opinion; — — —, prior based on an overview of trials; the range of equivalence is indicated by the two vertical dotted lines; the probabilities of falling below, within and above the range of equivalence are shown in the top right-hand corner, first for the clinical prior and then for the overview prior): (a) prior; (b) likelihood ($m = 59$, $x = -0.416$); (c) posterior; (d) predictive distribution

3.2.3. *Prior distributions*

Prior information regarding the effect of neutron therapy was available from two sources. First, the consensus belief of those interviewed gave a 28% belief that $\delta < 0$ (neutrons worse than photons) and 26% belief that $\delta > \delta_S$ (neutrons preferable to photons), with the remaining 46% lying in the range of equivalence. We have constructed, as an approximation, a normal prior distribution with mean log-hazard ratio $\delta_0 = 0.169$ and equivalent number of events $n_0 = 48$. We can think of this prior distribution as equivalent to having already observed 48 deaths, of which 22 were in the neutron therapy group and 26 in the photon therapy group. This prior is shown in Fig. 1(a) (full curve) and represents a group clinical opinion.

The second source of prior information was a statistical analysis of the combined results of previous randomized trials, conducted using different dosage regimens, which gave an estimated odds ratio for one-year mortality of 0.47 (95% interval 0.30–0.73) in favour of photon therapy. Relative to the assumed base-line survival of 50% under photons, this can be translated to an approximate normal prior, on the log-hazard-ratio scale, with $\delta_0 = -0.569$ and $n_0 = 138$. This prior distribution is particularly sceptical and, in view of the different regimens, might reasonably have been viewed with caution at the start of the trial.

3.2.4. *Likelihood*

The interim analysis, based on 59 observed deaths, showed an estimated hazard ratio of 0.66, with 95% interval (0.40, 1.10), comparing the death-rate in the conventional radiotherapy group with that in the neutron therapy group. Thus we have an estimated log(hazard ratio) $= -0.416$ with standard error 0.260, and using the notation of Section 3.1.1 for log-hazard ratios we obtain a likelihood with $x_m = -0.416$, $\sigma = 2$ and $m = 59$. This likelihood is shown in Fig. 1(b).

3.2.5. *Posterior distributions*

The probability that the effect of neutron therapy exceeds the minimal clinically worthwhile benefit δ_S is small (0.004) even under the more enthusiastic prior distribution. Moreover, the results of the current trial agree very closely with the combined results from previous trials. The data monitoring committee ratified the decision of the principal investigator to suspend entry of patients into the trial. The Bayesian analyses we have presented certainly support that decision.

4. PRIOR OPINION

4.1. *What Type of Prior?*

The term *prior* is used here to denote opinion based on evidence that is *external* to the trial: at the design stage this will be genuinely 'prior', in the temporal sense, but as the trial continues it is quite feasible that this prior will change because, for example, of the publication of related studies. Such changes should preferably be carried out by translating other studies into appropriate likelihoods.

When reporting studies we should acknowledge that different individuals or groups hold different prior beliefs. There is therefore no reason to select one particular specification, and instead we may consider a *community* of priors (Kass and Greenhouse, 1989) covering the perspectives of a range of individuals. This may

encompass a *reference* prior intended to add as little as possible to the data and a *clinical* prior expressing reasonable opinions held by individuals or derived from overviews (meta-analyses) of similar studies. However, it is also useful to develop 'off the shelf' priors corresponding to a formal expression of *sceptical* and *enthusiastic* belief—these may be thought to provide reasonable bounds to the community of priors.

4.1.1. *Reference priors*

Reference priors are supposed to represent minimal prior information, and for our simple normal likelihood they are obtained as the limit of the prior (2) as $n_0 \to 0$, leading to an improper uniform prior over the entire range of δ, and an identification of the posterior distribution with the normalized likelihood.

In some ways this is the most unrealistic of all possible priors. It represents, for example, a belief that it is equally likely that the relative risk associated with neutron therapy (Section 3.2) is above or below 10. However, it could be argued that such a prior is the least subjective and as such plays a useful role as a base-line against which to compare other more plausible priors.

4.1.2. *Clinical priors*

A clinical prior is intended to formalize opinion of well-informed specific individuals, often those taking part in the trial themselves. Deriving such a prior requires asking specific questions concerning a trial, and a variety of sources of evidence may be used as a basis for the opinion (see Sections 4.2 and 4.3).

4.1.3. *Sceptical priors*

One step towards incorporation of knowledge into a prior is an attempt to formalize the belief that large treatment differences are unlikely. Such an opinion could be represented by a symmetric prior with mean $\delta_0 = 0$ and suitably spread to include a range of plausible treatment differences. Kass and Greenhouse (1989) explicitly considered several such priors in their reanalysis of the extracorporeal membrane oxygenation data and introduced the concept of a 'cautious reasonable sceptic' who, if they held such a prior opinion, would consider it ethical to randomize patients since they did not hold a strong preference in favour of either treatment. This correspondence between belief in clinical superiority and the ethics of randomization is discussed further in Section 5. Sceptical priors may be particularly appropriate for regulatory authorities who are considering new drug applications.

A natural form of scepticism is to consider that the trial designers were optimistic in their expectation regarding the new treatment. This leads to a particularly simple form for the prior. Suppose that the trial has been designed with size α and power $1-\beta$ to detect an alternative hypothesis δ_A. Then we have the standard relationship

$$\sigma^2 \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta_A^2} = n \tag{4}$$

between the proposed sample size n and δ_A, where $z_\beta = \Phi(\beta)$. We express scepticism concerning δ by having a prior distribution which is normal with mean 0 and such that $p(\delta > \delta_A)$ is a small value γ. Such a distribution is shown as a full curve in Fig. 2 and has the property that

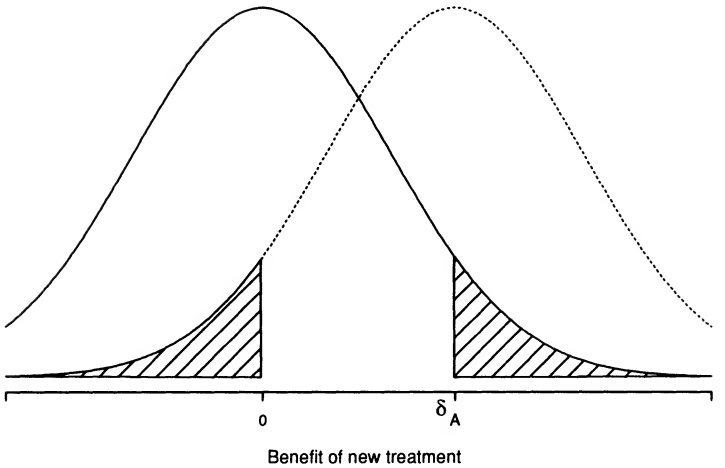

Fig. 2. Sceptical (———) and enthusiastic (· · · · ·) priors for a trial with alternative hypothesis δ_A: the sceptics' probability that the true difference is greater than δ_A is γ (shown shaded), which is also the enthusiasts' probability that the true difference is less than 0

$$\sigma \frac{z_{1-\gamma}}{\sqrt{n_0}} = \delta_A. \tag{5}$$

Equating δ_A in equations (4) and (5) gives

$$\frac{n_0}{n} = \left(\frac{z_{1-\gamma}}{z_{1-\alpha/2} + z_{1-\beta}} \right)^2.$$

Reasonable values might be $\alpha = 0.05$, $\beta = 0.1$ and $\gamma = 0.05$, which gives $n_0/n = 0.257$. Thus a formal expression of scepticism is obtained by assuming a prior equivalent to already having observed a quarter of the trial with a zero treatment difference. Such a prior will obviously provide a considerable handicap to the data showing a positive difference: the interesting consequence of this particular handicap will be explored in Section 8.1. Choosing n_0 in this manner only makes sense if the trial design is based on a realistic value for δ_A. If, as is often the case, δ_A is unrealistically large, then n_0 will no longer reflect the scepticism intended.

4.1.4. *Enthusiastic priors*

To counterbalance the sceptical prior discussed above, it is natural to introduce an enthusiastic prior representing individuals who are reluctant to stop when results supporting the null hypothesis are observed. Such a prior may have mean δ_A and the same precision as the sceptical prior, and hence provides a prior belief $p(\delta < 0) = \gamma$. This prior is shown in Fig. 2. In practice this formally derived prior is often similar to the clinical prior derived from investigators' opinions (see, for example, Section 4.3).

4.2. *Sources of Evidence for Clinical Priors*

4.2.1. *Evidence from other randomized trials*

If the clinical prior distribution is to represent current knowledge accurately, we should base it, where possible, on objective information. When the results of several

similar clinical trials are available, a statistical overview of those results can be used as the basis of a prior distribution. For example, the design of the beta-blocker heart attack trial (BHAT) was based on the results of five European trials of different beta-blocker drugs for the treatment of patients with recent acute myocardial infarction (BHAT Research Group, 1981). The investigators assumed that the drug propanalol would reduce mortality by 28% on the basis of these results. Although the researchers did not comment on the precision of this prior estimate, data from Yusuf *et al.* (1985) on these same five trials indicate that the standard error was approximately 10%. It is interesting that the final trial results showed a 28% reduction in mortality associated with propanalol.

Results from previous randomized trials should generally form the basis for a prior distribution but should not specify the distribution completely. As Kass and Greenhouse (1989) pointed out, a Bayesian who took this past evidence directly as his prior would be treating the historical and experimental subjects as exchangeable and essentially pooling the results. In the case of the BHAT, the evidence from previous studies might itself have been considered sufficiently strong to call into question the necessity for a further trial. However, it seems reasonable to combine previous data with some initial scepticism. Belief in heterogeneity of treatment effects across trials or patient subgroups, combined with reasonable scepticism, should suggest either shrinking the apparent treatment effect, expanding the variance or both. Random effects models in meta-analysis might be an appropriate tool, in which case the prior distribution would correspond to the predictive distribution for the effect in a new trial (Carlin, 1992).

4.2.2. *Evidence from non-randomized studies*

Often, however, no relevant randomized trial results are available, but non-randomized studies may have been conducted. The difficulty of constructing a prior distribution from the results of such studies is related to the assessment of the possible biases that can exist in such studies (Byar *et al.*, 1976). Such difficulty may often lead quite reasonably to a prior distribution with a large variance.

4.2.3. *Subjective clinical opinion*

We have emphasized that even when evidence in the form of randomized studies is available the prior distribution should be constructed by *using* that evidence but possibly applying a subjective adjustment. It may happen that there are no similar previous studies, in which case the distribution needs to be based on subjective judgment alone. One approach to eliciting opinion is to conduct individual interviews with clinicians who will participate in the trial.

We have reported this for several trials including the MRC trial of thiotepa for superficial bladder cancer (MRC Urological Working Party, 1985) and the European osteosarcoma intergroup trial of two adjuvant chemotherapy schedules for osteosarcoma (Spiegelhalter and Freedman, 1988). Two statisticians (Freedman and Spiegelhalter) interviewed individual clinicians in the manner described in Freedman and Spiegelhalter (1983). Interactive graphics could enhance this elicitation procedure, by providing fast quantitative feed-back and easy adjustment of distributional shape (Chaloner *et al.*, 1993).

Individual interviewing is time consuming and it may be preferable to use telephone or postal elicitation. A form for postal elicitation of prior opinion has been designed

TABLE 1
Summary of results of questionnaires completed by nine clinicians in the MRC CHART trial for head and neck cancer

Clinician	*Range of equivalence*	*Prior distribution (absolute advantage of CHART over standard therapy in 2-year % disease-free survival)*							
		−10 to −5	*−5 to 0*	*0–5*	*5–10*	*10–15*	*15–20*	*20–25*	*25–30*
1	5–10	10	30	50	10				
2	10–10	10	10	25	25	20	10		
3	5–10			40	40	15	5		
4	10–10			20	40	30	10		
5	10–15			20	20	60			
6	15–20	5	5	10	15	20	25	10	5
7	5–10				10	20	40	30	
8	20–20				10	20	40	20	10
9	10–15					10	50	30	10
Group mean	10–13	3	5	18	19	22	20	10	3

and is being used in several trials currently being co-ordinated from the MRC Cancer Trials Office. Below we report results from this questionnaire: similar techniques were used by Gore in trials of artificial surfactant (Ten Centre Study Group, 1987) and neutron therapy (Errington *et al.*, 1991).

4.3. *Example 2: Continuous Hyperfractionated Accelerated Radiotherapy Study in Head and Neck Cancer*

Our second example is a trial currently comparing standard therapy with continuous hyperfractionated accelerated radiotherapy, CHART (Parmar *et al.*, 1994). In designing the CHART study, base-line two-year disease-free survival under control therapy was estimated to be 45%. Nine individual prior distributions are summarized in Table 1, which presents the proportions of each distribution falling within intervals for treatment difference provided on the questionnaire.

As expected there are differences in the positions and shapes of the individual distributions. The opinions of clinicians 1 and 9 do not even intersect, whereas clinician 5 has a very confident (narrow) distribution, in contrast with clinician 6. How should we combine these individual distributions to arrive at a prior distribution for the group? Many methods have been proposed (Genest and Zidek, 1986). The two simplest methods are arithmetic and logarithmic pooling, corresponding to taking the arithmetic and (normalized) geometric mean respectively within each column of Table 1. The former takes the opinions as data and averages, whereas the latter takes the opinions as *representing* data and pools those implicit data. Our strong preference is for arithmetic pooling, to obtain an estimated opinion of a typical participating clinician.

Arithmetic pooling for the CHART study gives a median of 11% improvement, with a prior probability 0.08 that there is a deterioration (δ_0). Assuming a 45% two-year disease-free survival rate under standard therapy, and transforming to a log-hazard-ratio scale this corresponds to $p_0(\delta)$ having a median of $\log(\log 0.45/\log 0.56) = 0.32$, with $p_0(\delta < 0) = 0.08$. Assuming a normal prior distribution with these characteristics gives parameters $\delta_0 = 0.32$ and $n_0 = 77$. This

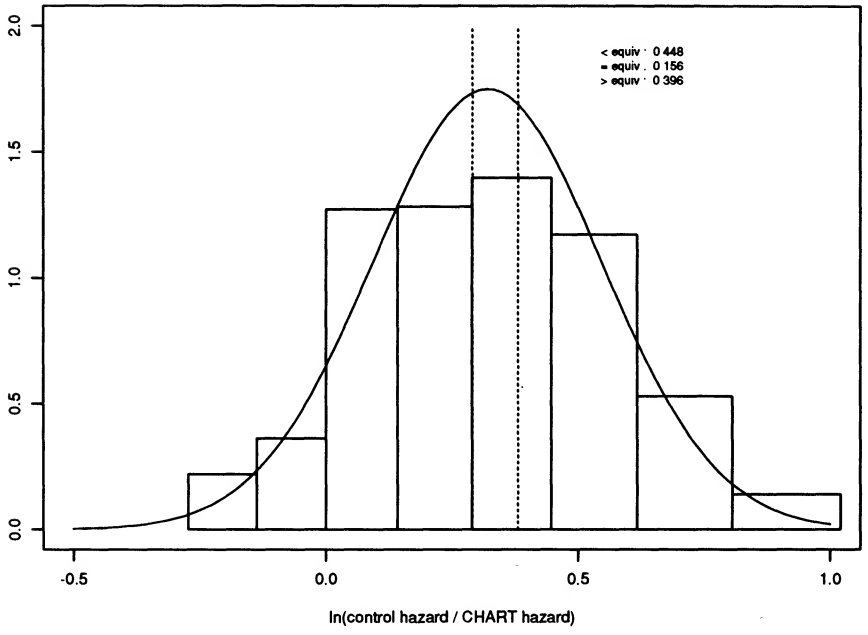

Fig. 3. Prior for log(control hazard/CHART hazard) derived from elicited subjective opinions

prior distribution is shown in Fig. 3, superimposed on a histogram derived from the group distribution shown in Table 1—it can be seen that the normal approximation is adequate, and we have generally found this to be the case. The mean of the ranges of equivalence that were elicited from the same nine clinicians, (10%, 13%), is shown transformed to (0.29, 0.38) on the δ-scale. The ethical basis for the randomization is clear, since the prior probabilities of being on either side of the range of equivalence are almost equal (0.45, 0.40). Such a juxtaposition of prior opinion and range of equivalence formalizes a grouped version of the 'uncertainty principle' (see, for example, Byar *et al.* (1990)), under which it is considered ethical to randomize a patient if the clinician has reasonable uncertainty about which is their most appropriate treatment.

Section 6.2 contains an example of the use of this distribution in pretrial power calculations.

5. MONITORING OF TRIALS

5.1. *Bayesian Monitoring Criteria*

Ethical considerations place an obligation on organizers to take into account all sources of evidence when considering the continuation of a trial (Pocock and Hughes, 1989), not purely the data from the trial itself. A rational way to do this is to use the trial data to provide the likelihood and have all external evidence concerning the effect of treatment on the main clinical outcome represented by a prior distribution. Their combination into a posterior distribution yields the intervals that can be contrasted with the range of equivalence, which combines evidence regarding the relative differences between the treatments in toxicity, cost and inconvenience.

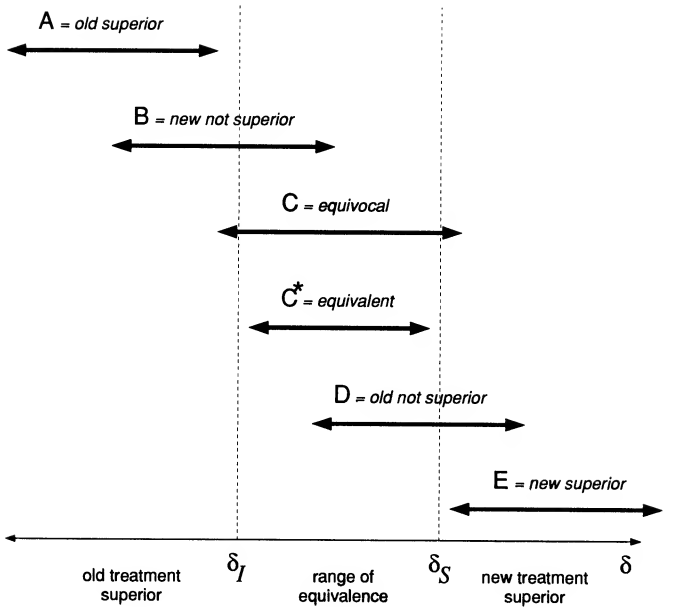

Fig. 4. Possible situations at any point in a trial's progress, derived from superimposing an interval estimate (say 95%) on the range of equivalence (see the text for the relationship between situations A, B, C, C*, D and E and monitoring trials)

In this way internal and external evidence on the relative benefits and risks of the treatments are being combined and weighted quantitatively.

Our proposal for monitoring simply extends the preceding discussion on ethical randomization, the judgment being based on the posterior rather than the prior distribution. If clinicians are essentially certain that one treatment is clinically superior, then they should not randomize further patients; if they are almost certain that one treatment is not superior, and may be inferior, then they should carefully consider stopping randomization. This approach has already been illustrated by the neutron therapy example in Section 3.2, but is illustrated in generality in Fig. 4.

In Fig. 4 are shown six hypothetical credible intervals for the treatment difference (i.e. intervals containing a nominal percentage, say 95%, of the posterior distribution) in relation to the range of equivalence. Each interval corresponds to a different statement that may be made regarding the comparison of the treatments. If the interim analysis yields (credible) intervals which have the same form as A or E, then a recommendation to terminate the trial would generally be indicated (Armitage, 1989). Intervals of the form of B or D may also warrant termination, depending on the nature of the two treatments. Errington *et al.* (1991) described a trial that was stopped in situation B. Interval C* is of particular interest within equivalence studies. Meier (1975) has discussed similar procedures for monitoring trials.

From a posterior distribution it is straightforward to calculate the probability content of the three crucial areas: below, within and above the range of equivalence. Using the intervals B or D focuses attention on whether the area above or the area below the range of equivalence has become a sufficiently small value ϵ to consider stopping the trial. We are reluctant to express a firm opinion about what this critical size ϵ

should be, since the choice should, in principle anyway, be made from decision theoretic considerations of expected utility. But we have already said that this is unrealistic, and so we are left with the conventional bench-marks such as 2.5% and 5%. However, the selected values can be informally varied according to the perceived importance of the conclusion: for trials of therapy likely to have wide implications for health care, we recommend a stringent criterion. Mehta and Cain (1984) have described essentially this procedure for use in phase 2 studies.

5.2. *Which Prior to Use?*

The preceding discussion has presupposed that a single posterior distribution is available. However, we have already argued that a community of prior distributions should be considered, and hence monitoring will give rise to a range of possible posterior distributions, possibly based on reference, clinical, sceptical and enthusiastic priors.

Although tail areas based on a clinical prior may best reflect the opinions of the trial organizers and participants, they should keep in mind the need for a trial to provide convincing evidence to a spectrum of reasonable opinion. This fits with the recommendation within, for example, the AXIS study (United Kingdom Coordinating Committee on Cancer Research, 1989) that the data monitoring committee should only alert the steering committee if there is

> '**both** (a) "proof beyond reasonable doubt" that for all, or for some, types of patient one particular treatment is clearly indicated . . ., **and** (b) evidence that might reasonably be expected to influence the patient management of many clinicians who are already aware of the results of other main studies'.

Therefore we recommend that when considering stopping a trial on the basis of results that indicate an effect, either beneficial or harmful, of a new treatment, a *sceptical* prior should be examined; when confronted with data that indicate little or no effect of the new treatment an *enthusiastic* prior should be considered (see Section 7 for examples of this approach). In this way the trial will only stop early if sufficient evidence has been provided to counterbalance the prior opinions of someone who would doubt the observed results. In part, the prior is set up as representing an adversary who will need to be disillusioned by the data to stop further experimentation—for a formalization of this process see Lindley and Singpurwalla (1991).

6. PREDICTIONS

6.1. *Making Predictions*

A major strength of the Bayesian approach is the ease of making predictions concerning events of interest. Suppose that we have observed m observations and are interested in the possible consequences of continuing the trial for a further n observations. If we denote our future statistic by X_n, then it has predictive distribution

$$p_m(X_n) = \int p(X_n|\delta)\, p_m(\delta)\, \mathrm{d}\delta$$

and it is straightforward to show that given the posterior distribution (3) we obtain

$$p_m(X_n) = \phi\left\{X_n \middle| \frac{n_0\delta_0 + mx_m}{n_0+m},\ \sigma^2\left(\frac{1}{n_0+m} + \frac{1}{n}\right)\right\}. \tag{6}$$

As a special case, if we have not observed any data so far ($m=0$) we have that

$$p_0(X_n)=\phi\left\{X_n\,|\,\delta_0,\,\sigma^2\left(\frac{1}{n_0}+\frac{1}{n}\right)\right\}. \tag{7}$$

The use of this expression in sample size determination is described in Section 6.2. Section 6.3 concerns a measure of conflict between the prior distribution and the likelihood, and interim predictions are discussed in Section 6.4.

6.2. *Pretrial Predictions*

Suppose that at the start of a trial ($m=0$) we wish to assess the chance that the trial will arrive at a firm positive conclusion, i.e. the final interval will exclude a value δ_I in favour of the new treatment. Here we assume that the target final interval will be based on the reference posterior, equivalent to the normalized likelihood. As outlined in Section 2.2, the standard recommended procedure in this context is to select an alternative hypothesis δ_A and to use the resulting 'power' calculation as a basis for selecting a suitable sample size. We have argued that this alternative hypothesis should reflect realistic expectation, and hence it is a natural extension to acknowledge the uncertainty in the anticipated treatment difference expressed in the full prior distribution. Power over the plausible range of δ should be calculated, and a summary measure might be the expected power.

This can be derived as follows. The critical value of X_n is

$$X_n>\delta_I-\frac{1}{\sqrt{n}}z_\epsilon\sigma=x^*$$

which from equation (7) will occur with probability

$$p_0(X_n>x^*)=\Phi\left\{z_0(\delta_I)\sqrt{\left(\frac{n}{n_0+n}\right)}+z_\epsilon\sqrt{\left(\frac{n_0}{n_0+n}\right)}\right\} \tag{8}$$

where $z_0(\delta_I)=(\delta_0-\delta_I)\sqrt{n_0}/\sigma$ is the standardized distance of the prior mean from δ_I. As $n_0\to\infty$ we obtain the classical power curve

$$\Phi\left\{(\delta_0-\delta_I)\frac{\sqrt{n}}{\sigma}+z_\epsilon\right\}.$$

This use of the full prior distribution for power calculations leads to what has been termed predictive power (Spiegelhalter *et al.*, 1986), expected power (Brown *et al.*, 1987) or the strength (Crook and Good, 1982) of the study. Spiegelhalter and Freedman (1986) provided a detailed example of the predictive power of a study using subjective prior opinions from a group of urological surgeons to form the prior distribution, whereas Moussa (1989) discussed the use of predictive power with group sequential designs based on ranges of equivalence. Brown *et al.* (1987) have also suggested the use of prior information in power calculations, but conditioning on $\delta>\delta_S$, i.e. the predictive probability of concluding the new treatment is superior, given that it truly is superior.

6.2.1. *Example 2 (continued)*

The CHART trial was originally designed with an alternative hypothesis of an absolute 15% improvement of CHART over the 45% two-year disease-free survival estimated for the standard therapy, equivalent to $\delta_A=\log(\log 0.45/\log 0.60)=0.45$.

With a two-sided 5% test the trial had 90% power for the 218 events expected. Using the prior shown in Fig. 3, the predictive probability that a final 95% interval will exclude 0 is 83%, 62% that it will be in favour of CHART and 21% that it will be in favour of control. However, the ranges of equivalence provided by the clinicians should warn us that simply proving 'significantly better than no difference' may not be sufficient to change clinical practice. In fact, at the first interim analysis the data monitoring committee extended recruitment to expect 327 events. This decision was made independently of knowledge of interim results: using the same prior opinion the predictive probability of concluding in favour of CHART is now 68%, and in favour of control 25%.

Spiegelhalter and Freedman (1988) noted that the predictive power (8) could be expressed as the power against a fixed alternative δ_u, where

$$\delta_u = \delta_{0.5}\left\{1 - \sqrt{\left(\frac{n_0}{n_0+n}\right)}\right\} + \delta_0\sqrt{\left(\frac{n_0}{n_0+n}\right)},$$

a weighted average of the prior mean and the point of 50% power ($\delta_{0.5}$ is the value which, if observed, would just lead to rejection of the null hypothesis). Thus the predictive power will always lie between 0.5 and the power calculated at the prior mean. Predictive power is a useful addition to simple power calculations for a fixed alternative hypothesis and may be used to compare different sample sizes and designs for a proposed trial, and as an aid to judge competing trials considering different treatments.

6.3. *Checking Prior–Data Compatibility*

We have suggested that, before a trial starts, there should be extremely careful consideration of what the treatment difference is likely to be, and that this should be used as an important component for trial design. However, it may easily happen that these initial judgments are misguided and the data do not fit the stated expectations. Far from invalidating the Bayesian approach, such a conflict between prior and data only emphasizes the importance of pretrial elicitation of belief: having these opinions explicitly recorded will help a monitoring committee to focus on the difference between expected and actual results. It is useful to have a formal means of assessing the confict between the prior expectations and observed data, although the precise action to be taken in the face of considerable conflict will depend on the circumstances.

Box (1980) suggested using the prior to derive a predictive distribution, and then to calculate the chance of a result with lower predictive ordinate than that actually observed. In our context we would use the form of equation (7) but substituting m for n, to give a pretrial predictive distribution

$$p_0(X_m) = \phi\left\{X_m \mid \delta_0, \sigma^2\left(\frac{1}{n_0} + \frac{1}{m}\right)\right\}. \tag{9}$$

Given observed x_m, Box's generalized significance test is given by

$$p_0\{p_0(X_m) \leqslant p_0(x_m)\} = 2\min\{P_0(x_m),\ 1 - P_0(x_m)\},$$

since the predictive distribution is unimodal.

For example, in Fig. 1 the predictive distributions $p_0(X_m)$ are drawn for the two priors, and the dotted line indicates the sample log-hazard ratio. Under the clinicians' prior there was only a $2 \times 0.07 = 0.14$ probability of observing such an unlikely statistic, whereas under the overview prior the observed statistic was quite unsurprising.

We deliberately do not specify the action to take on observing substantial conflict between prior and data. As with any global significance test, this must depend on individual circumstances.

6.4. *Interim Predictions*

A frequent question asked by investigators at interim analyses is: given the data so far, what is the chance of obtaining a 'significant' result? Many researchers have considered an appropriate response under the heading of stochastic curtailment (Halperin *et al.*, 1982; Spiegelhalter *et al.*, 1986; Choi and Pepple, 1989; Ware *et al.*, 1985), and this work is reviewed by Jennison and Turnbull (1990). Within a Bayesian framework the question arises whether the prior opinion is included in making the predictions and in the analysis (the fully Bayesian case), just in making the predictions (the mixed case) or ignored completely (the likelihood case). The appropriate formulae for each of these options are given in Appendix B. We emphasize that this analysis assumes that future summary statistics are drawn from exactly the same distribution as for those statistics observed so far. However, often the predictions are made over an extension of the follow-up of recruited patients, rather than for extending the numbers of new patients. Hence the adequacy of the predictions depends on a strong, and possibly unwarranted, assumption of proportional hazards.

An important issue concerns the use of such predictions. We follow Armitage (1989) in claiming that the decision on whether there is sufficient evidence to warrant stopping a study should be based on the current opinion summarized by the posterior distribution. Predictions concerning the consequences of continuing, although superficially attractive, give an undue weight to achieving 'significance' and are of secondary importance.

7. EXAMPLES OF BAYESIAN ANALYSIS

Most published Bayesian analyses are re-examinations of trial results originally published after using conventional techniques. Examples include the APSAC intervention mortality study trial of APSAC against a placebo (AIMS Trial Study Group, 1988) for which Pocock and Hughes (1989) showed the shrinkage of the observed treatment effect brought about by considering a plausible prior distribution; they also considered the consequences of a range of priors. Pocock and Spiegelhalter (1992) showed a similar use of shrinkage after an extreme result was reported in a trial of home thrombolytic treatment. A trial in colorectal cancer originally reported by Poon *et al.* (1989) has been reanalysed and discussed by Dixon and Simon (1992), Freedman and Spiegelhalter (1992) and Greenhouse (1992). Here we consider two further examples in detail.

7.1. *Example 3: Levamisole and 5-fluorouracil in Bowel Cancer*

7.1.1. *The trial*

Moertel *et al.* (1990) have reported results from a randomized clinical trial investigating the effect of the drug levamisole (LEV) alone or in combination with

5-fluorouracil (5-FU) for patients with resected cancer of the colon or rectum. Patients entering the trial were allocated one of the three treatments LEV, LEV + 5-FU or control. The main outcome measure of treatment was the duration of survival. Moertel *et al.* (1990) planned their study with an alternative of $\delta_A = \log 1.35 = 0.30$, and to have 90% power for a one-sided test with 5% size required 380 deaths.

7.1.2. *Range of equivalence*

In this context Fleming and Watelet (1989) suggested a range of equivalence, on a control-to-treatment relative hazard scale, of (1.0, 1.33), so that on a log-hazard-ratio scale $\delta_I = 0$ ($\log 1.0$) and $\delta_S = 0.29$ ($\log 1.33$), i.e. the control treatment would be clinically preferable provided that it had no excess mortality, whereas the new treatment, with its 'inconvenience, toxicity and expense', would only be clinically superior if it reduced the hazard by at least 25% ($1 - 1/1.33$).

7.1.3. *Prior distributions*

Since we are conducting a retrospective analysis of this trial it is reasonable to investigate sceptical and enthusiastic priors: we note that in this study, as with many others, the range of equivalence (δ_I, δ_S) and the interval (0, δ_A) essentially coincide. We adopt a sceptical prior with mean 0 and probability 0.05 of exceeding $\delta_A = 0.30$, and assuming $\sigma = 2$ gives a prior with $\delta_0 = 0$ and $n_0 = 120$; this prior is quite reasonable in view of the implausibility of dramatic gains in cancer adjuvant therapy. The enthusiastic prior has the same precision and mean 0.30. These priors are shown in Fig. 5(a).

7.1.4. *Likelihood*

From Moertel *et al.* (1990), $x_m = 0.04$ and $m = 223$ for LEV *versus* control, and $x_m = 0.40$ and $m = 192$ for LEV + 5-FU *versus* control. The likelihood for LEV + 5-FU *versus* control is shown in Fig. 5(b)—the reference posterior (normalized likelihood) shows convincing evidence that $\delta > 0$ (prob = 0.003), but moderate evidence (prob = 0.78) that the treatment is clinically superior. The predictive distribution shows that there was only 4% chance of observing such a high result given the sceptical prior.

7.1.5. *Posterior distributions*

The sceptical and enthusiastic posterior distributions for LEV + 5-FU *versus* control are shown in Fig. 5(c). Since the result is positive we should concentrate on the sceptical prior to see whether the interim data are sufficient to convince someone holding this opinion. The sceptical posterior distribution has mean 0.25 with standard deviation 0.11, corresponding to an estimate of 1.28 and 95% interval (1.03, 1.60) for the hazard ratio. There is a small posterior probability (0.015) of the treatment difference being less than 0, i.e. of the control treatment being superior to LEV + 5-FU, whereas the posterior mean is within the range of equivalence. We might therefore argue that the result has not yet achieved 'practical significance' to a reasonable sceptic, and hence that the trial might have continued. This opinion is possibly supported by the fact that this treatment has not become accepted in Europe where trials are still continuing.

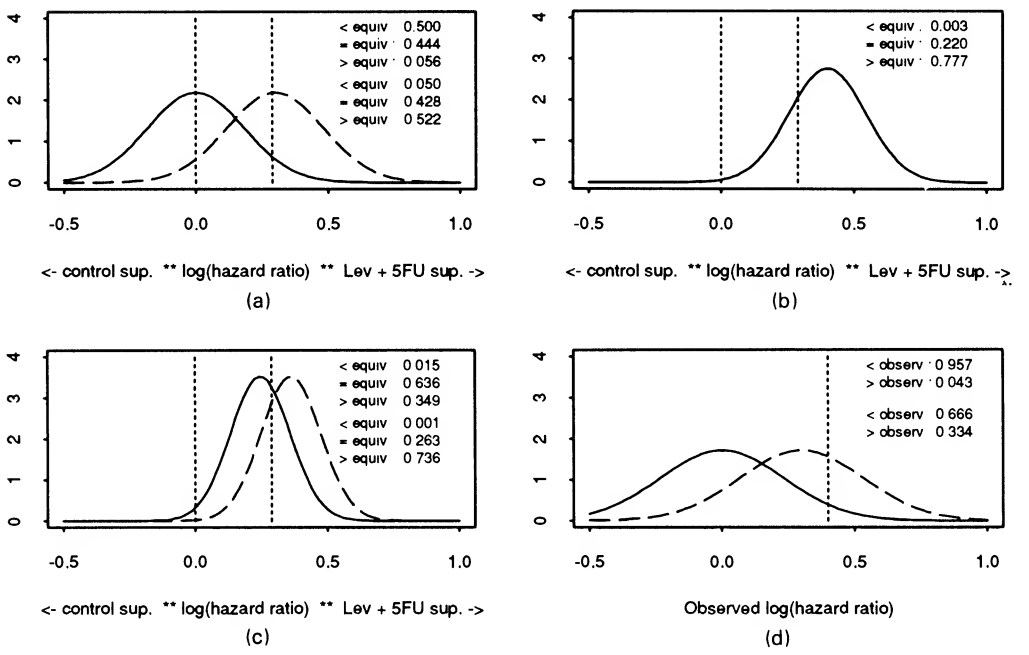

Fig. 5. Prior, likelihood and posterior distributions for LEV + 5-FU *versus* control (———, sceptical prior and posterior; — — —, enthusiastic prior and posterior; the probabilities of falling below, within and above the range of equivalence are shown in the top right-hand corner, first for the sceptical prior and then for the enthusiastic prior): (a) prior; (b) likelihood ($m = 192$; $x = 0.4$); (c) posterior; (d) predictive distribution

Results are also available from an earlier study (Laurie *et al.*, 1989) which showed an apparent reduction in the death-rate of patients receiving LEV + 5-FU, with an estimated log-hazard ratio of 0.14 with standard error 0.17. If we include these data in the analysis under the sceptical prior our hazard ratio estimate is changed to 1.25 with 95% interval (1.04, 1.49). Including these data thus increases certainty that $\delta > 0$ (prob = 0.007) but decreases the certainty that the treatment is clinically superior, $p(\delta > 0.29 \mid \text{combined evidence}) = 0.22$, and so does not change the interpretation given above.

7.1.6. *Interim predictions*

The investigators' original plan was to include approximately 188 further deaths in the comparison of control and LEV + 5-FU. Table 2 shows the predictive probability of different conclusions being drawn at the end of this period, calculated under the three assumptions

(a) a fully Bayesian analysis using our sceptical prior,
(b) that the prior is used for predictions but excluded from the final analysis and
(c) that the predictions and analysis are based solely on the data (equivalent to using the reference prior throughout).

TABLE 2
Predictive probability of the final 99% interval for the hazard of control relative to LEV+5-FU having different positions relative to the range of equivalence, after observation of a further 188 events†

Position of 99% interval	*Results for the following modes of prediction:*		
	Bayesian	*Mixed*	*Likelihood*
A: old superior	0.000	0.000	0.000
B: new not superior	0.004	0.001	0.000
C: equivocal	0.407	0.256	0.091
D: old not superior	0.590	0.737	0.845
E: new superior	0.000	0.006	0.064

†For illustration of A, B, C, D and E see Fig. 3.

Ignoring the prior completely leads to a final interval that is very likely to exclude $\delta = 0$ (D or E), which might reinforce the decision to stop the trial. However, if the prior is used for prediction then the chance that a final 99% interval includes 0 rises to 0.26, whereas the chance that the sceptical posterior interval includes 0 is 0.41. Thus the eventual result of the study is not a foregone conclusion (we emphasize the strong dependence on the proportional hazards assumption for this analysis, as observing future events depends on extending the follow-up of patients already in the trial rather than new recruitment).

7.2. *Example 4: Medical Research Council Misonidazole Trials*

7.2.1. *The trial*

In the summer of 1979 the MRC decided to investigate in three randomized trials the effect of the drug misonidazole (MISO) that was thought to enhance the effects of radiotherapy. The trials involved the treatment of three different cancers (brain, head and neck, and cervix cancer), for each of which X-ray therapy (XRT) was the main therapy. Each trial compared the outcomes of patients treated by XRT plus MISO with those treated by XRT alone. We focus here on the head and neck cancer trial. For this trial the primary end point was the duration of control of the primary tumour. Patients began to enter this trial in early 1979. The design required 292 events in two studies, giving 90% power to obtain a significant result at the two-sided 5% level under the alternative hypothesis $\delta_A = 0.38$.

7.2.2. *Range of equivalence*

After about 1 year's experience it became apparent that MISO carried a higher risk of serious side-effects, mostly peripheral neuropathies, than had been thought previously. Such neuropathy took its most severe form in complete numbness or cramp of the limbs and extremities. The problem occurred more frequently in the head and neck, and cervix cancer patients than in the brain cancer patients. Clinicians began to express concern and wished to know whether the early results indicated any benefits that would offset the increased toxicity due to MISO. In discussion with clinicians, minimum clinically worthwhile treatment differences were defined that could act as

decision points in interim analyses. For head and neck cancer this was an increase in the two-year primary control from 25% to 40%, i.e. a hazard ratio of 1.513; we may think of $\log 1.513 = 0.414$ as the upper limit δ_S of the range of equivalence on a log-hazard-ratio scale.

7.2.3. *Prior distributions*

We consider three different prior distributions: a reference prior leading to an analysis based on the likelihood alone, a sceptical prior with mean 0 and implicit sample size equal to a quarter of the planned accrual in the trial ($n_0 = 73$) (i.e. a 'handicap' of 0.25 in the language of Section 8.1) and a more enthusiastic prior with the same standard deviation as the sceptical prior but centred on the alternative hypothesis $\delta_A = 0.38$.

7.2.4. *Likelihood*

Interim analyses were conducted approximately once per year. The third such analysis was completed in September 1981 based on 108 events, which gave a hazard ratio estimate of 0.90 (95% interval 0.62–1.31), corresponding to a log-hazard-ratio estimate of -0.108 with standard error 0.193.

7.2.5. *Posterior distributions*

Table 3 shows the consequences of assuming each type of prior opinion. The hazard ratio estimate for the sceptical posterior is 0.94 (95% interval 0.70–1.26), whereas for the enthusiastic posterior we obtain 0.897 (95% interval 0.82–1.46). These analyses indicate that, even with the more enthusiastic prior opinion, the chances are very small that MISO leads to benefits that the clinicians considered to be sufficient to justify the risk of toxicity.

Following the September 1981 interim analysis the MRC decided to terminate this trial (and the other two trials). The analyses presented here support that decision, although we note that Bayesian methods were not used at the time. Final reports of these trials may be found in references MRC Working Party on Misonidazole in Gliomas (1982), MRC Working Party on Misonidazole in Head and Neck Cancer (1983) and MRC Working Party on Misonidazole in Cancer of the Cervix (1983).

TABLE 3

Analysis of head and neck trial of MISO therapy with different prior assumptions for δ = log(control hazard/treatment hazard)†

Prior	*Prior for δ mean*	*Prior standard deviation*	*Posterior mean*	*Posterior standard deviation*	$p_m(\delta \geq \delta_S)$
Reference	0.0	∞	−0.108	0.193	0.004
Sceptical	0.0	0.234	−0.064	0.149	0.001
Enthusiastic	0.380	0.234	0.088	0.149	0.016

†For the definitions of the various priors see the text.

8. PRACTICAL CONSIDERATIONS

The long running debate between frequentist and Bayesian approaches to clinical trials (Cornfield, 1966; Berry, 1987; Whitehead, 1992) has largely been ideological but in this paper we are attempting to be pragmatic. Since the differences between the two approaches are greatest in the monitoring and reporting of clinical trial results, we concentrate on those issues here.

8.1. *What about Type I Error (α-levels)?*

Frequentist sequential schemes aim to control the overall type I error α. An important consequence is conservatism, particularly at the early stages of a trial, where the degree of conservatism depends on the stopping rule chosen. Some have advocated that this choice should be determined by the plausible size of the treatment difference: for example, a Pocock-type rule (Pocock, 1977) may be suitable when large differences are more plausible, whereas when large differences are unlikely an O'Brien–Fleming-type rule (O'Brien and Fleming, 1979) may be appropriate (Armitage, 1985). Freedman and Spiegelhalter (1989) showed that explicit representation of such scepticism as a prior distribution centred on zero difference leads to conservative behaviour similar to frequentist group sequential schemes. (We are assuming here that the range of equivalence is itself centred on 0 and is very narrow. If its centre is non-zero, and the prior distribution has mean equal to this non-zero value, a simple transformation of the scale will result in a monitoring scheme with the same statistical properties.)

If we are very concerned about sampling properties of trial monitoring procedures it is possible (Grossman *et al.*, 1994) to consider the frequentist properties of a Bayesian monitoring procedure in which the trial is terminated at an analysis if a Bayesian interval excludes 0. Table 4 shows the handicaps, expressed as a fraction of the total sample size, which fix the type I error to be 0.05 (when monitoring with a 95% posterior interval) and 0.01 (when using a 99% interval). As an example, if there are five planned

TABLE 4

Fraction of study size (the handicap) which is assumed to underlie a prior distribution with mean 0 that produces Bayesian stopping rules based on posterior $100(1-\alpha)\%$ intervals which have total type I error α

No. of analyses	*Handicaps giving the following values of α:*	
	$\alpha = 0.05$	$\alpha = 0.01$
1	0	0
2	0.16	0.11
3	0.22	0.15
4	0.25	0.17
5	0.27	0.18
6	0.29	0.20
7	0.30	0.21
8	0.32	0.22
9	0.33	0.22
10	0.33	0.23

analyses and we wish to preserve the overall type I error at 0.05, we could use a 95% posterior interval assuming a prior equivalent to having already conducted a trial in which 27% of the currently planned total sample size had been entered and no treatment difference was observed. This particular degree of sceptical prior was derived by other means in Section 4.1.2. We note that the handicap is fairly insensitive to the exact number of looks, so this Bayesian monitoring scheme for a typical trial with 3–5 analyses will have type I error roughly corresponding to the associated critical tail area of the posterior distribution. Grossman *et al.* (1994) have shown that such Bayesian–frequentist rules have good properties in terms of average sample number.

8.2. *Can We Let People Stop When They Want?*

An extreme example of optional stopping is provided by the theoretical possibility of 'sampling to a foregone conclusion', in which asymptotically we are guaranteed at some point to obtain a significant result even if the null hypothesis is true (McPherson, 1974). Cornfield (1966) argued that, if you are worried by this, it must reflect consideration of the null hypothesis as having a distinct probability of being true. It follows that we should put a lump of probability (however small) on the null hypothesis, and then the phenomenon will not occur.

Pocock and Hughes (1989) 'feel that control of the overall type I error is a vital aid to restricting the flood of false positives in the medical literature'. From a Bayesian perspective control of the type I error is not central to making valid inferences and we would not be particularly concerned with using a sequential scheme with a type I error that is not exactly controlled at a specified level. In fact, if it were not for reporting bias, we would see no objection to optional stopping. For example, if we were carrying out an overview of all randomized trials of a given question, and we knew the current results of all trials taking place, the type I error of any individual trial would be of little interest.

However, there are two pragmatic reasons why we might wish to employ monitoring schemes that have low type I error. First, reporting bias makes it more likely that trials with 'significant' results are published in journals with high reputations, and also more likely that 'non-significant' trials remain unpublished (Easterbrook *et al.*, 1991). Hence a high type I error rate would lead to greater numbers of published false positive results, some of which may remain uncountered by published negative reports. Second, a monitoring scheme that does not protect against early stopping is likely to lead to earlier publication of a false positive result that will only be countered several years later by a negative trial that has gone its full course. Similar arguments hold for trials which are conducted in the pharmaceutical industry where each trial is not necessarily published but is included in a submission to obtain a licence for a new treatment. Thus, from a practical point of view, we are largely in agreement with Pocock and Hughes, although it should be emphasized that sequential analysis was not intended as a counterbalance to publication bias.

However, it must be understood that the two concerns above arise from a background in which many, perhaps most, new experimental treatments are found to be ineffective. It is this same context that leads us to recommend the use of sceptical prior distributions, which both reflect the background information and provide a conservative approach to early terminations of trials. In a research environment where new treatments are usually found to be effective, the conservatism implicit in group

sequential schemes and explicit in a sceptical prior distribution could be unattractive. Likewise, in exceptional circumstances, there may be such compelling evidence for a new treatment that the background scepticism is overcome and a less conservative approach to monitoring may reasonably be adopted. Such a situation may arise in a confirmatory trial, i.e. a trial that seeks to reproduce positive results in one or more previously conducted randomized studies.

8.3. *Which Prior Should We Use and When?*

An issue that is central to the application of Bayesian methods is the choice of prior distribution. Clearly the 'stronger' (i.e. the more informative) the prior distribution, the more influence it has on the analysis. In particular, in monitoring trials, the prior distribution will have considerable influence in the early stages of a trial when relatively little data are available. Rosenbaum and Rubin (1984) have also pointed out that with optional stopping the coverage properties of the final Bayesian intervals are sensitive to the chosen prior distribution.

These considerations lead us to recommend that a community of priors should be used in monitoring and reporting. These should include, for monitoring, a reference, sceptical and enthusiastic prior. The sceptical prior should have mean 0 and precision discussed in Section 4.1.3: the effect will be to put a brake on early stopping in favour of a treatment difference. The enthusiastic prior might be an assessed clinical prior, or a prior from a meta-analysis or possibly a prior with mean δ_A, the alternative hypothesis, and precision the same as the sceptical prior. Such a prior tends to act against early stopping in favour of no difference or in favour of the control. Other options may be discussed when reporting the final results.

8.4. *How Can Subjective Judgments be Allowed into Reporting Trial Results?*

The methods that we have introduced are to aid interpretation of the trial results. Thus we suggest explicit separation of the results of a trial from their interpretation. In reporting, we suggest that the results are placed in their own traditional section, presenting means, standard errors, survival curves etc., whereas the Bayesian methods would be included in an additional formal section on 'interpretation'. In this interpretation section it would be important to describe the source of any prior distributions used and to show the sensitivity of any conclusions to a community of prior distributions such as that discussed earlier. This section would include interval estimates based on posterior distributions and place particular emphasis on the posterior probabilities of the treatment difference lying below, within and above the range of equivalence. This answers the crucial question: what is the chance, given evidence both internal and external to this trial, that a specific treatment is clinically superior?

8.5. *How Can Bayesian Calculations be Performed?*

One of the main blocks against the use of Bayesian techniques has been the lack of suitable software. The calculations and plots for this paper were all carried out by using a set of functions written in S (Becker *et al.*, 1990). These functions are available from the first author on request. The software considerably reduces the time and effort required to perform Bayesian analyses, to the extent that there

is only a slight addition to the time required for the standard analysis. It also provides the best means of presenting the results—graphically.

8.6. *Conclusions*

We have suggested a pragmatic set of tools to help in the design, monitoring and reporting of clinical trials. In deciding whether to start a trial, how large to make it, whether to stop and whether the results of the trial and other studies are together convincing to practising clinicians, a team will generally make informal assessments of many factors, e.g. the likely variability, the plausible benefit and what is sufficient evidence to render randomization inappropriate. All that we are proposing is that such assessments are made more *formal*, to clarify the issues and to provide a rational and explicit basis for both discussion of the issues and for decision-making. In our experience such tools can only help in the communication between statisticians and their clinical colleagues.

ACKNOWLEDGEMENTS

We thank the CHART steering committee for providing the data for example 2 and numerous referees for helpful comments.

APPENDIX A: MORE GENERAL FAMILY OF PRIOR DISTRIBUTIONS

In Section 3 we introduced a simple Gaussian prior distribution and described how a flexible family of distributions may be obtained by allowing a two-component mixture of Gaussian, truncated Gaussian and degenerate distributions (known as 'lump' priors). Applications of such priors are contained in Sasahara *et al.* (1973) and Freedman and Spiegelhalter (1992). Here we provide the technical details for using this prior family for prior-to-posterior analysis, prior–likelihood comparison and those predictive statements that are available in closed form, assuming a normal likelihood given in equation (1).

A.1. *Lump Priors*

Let $p_0^{\text{lump}}(\delta)$ denote a prior with a degenerate mass on $\delta = \delta_0$. Observing data x_m leads to a posterior distribution that is unchanged from the prior. The predictive ordinate of the data may be obtained by letting $n_0 \to \infty$ in expression (9), giving

$$p_0^{\text{lump}}(x_m) = \phi(x_m \mid \delta_0,\ \sigma^2/m).$$

A.2. *Truncated Prior*

We suppose a Gaussian density that has been truncated below at δ_L and above at δ_U. The prior density is denoted

$$p_0^{\text{trunc}}(\delta) = \phi(\delta \mid \delta_0,\ \sigma^2/n_0,\ \delta_L,\ \delta_U)$$

and is given by

$$p_0^{\text{trunc}}(\delta) = \begin{cases} p_0(\delta)/[\Phi\{z_0(\delta_L)\} - \Phi\{z_0(\delta_U)\}] & \delta_L < \delta < \delta_U, \\ 0 \quad \text{otherwise} \end{cases}$$

where $p_0(\delta)$ is from expression (2) and $z_0(\delta) = (\delta_0 - \delta)\sqrt{n_0}/\sigma$ (see equation (8)). Combined with likelihood (1) we obtain the posterior distribution

$$p_m^{\text{trunc}}(\delta) = \begin{cases} p_m(\delta)/[\Phi\{z_m(\delta_L)\} - \Phi\{z_m(\delta_U)\}] & \delta_L < \delta < \delta_U, \\ 0, \quad \text{otherwise} \end{cases}$$

where $p_m(\delta)$ is from expression (3) and

$$z_m(\delta) = \left(\frac{n_0\delta_0 + mx_m}{n_0 + m} - \delta \right) \frac{\sqrt{(n_0 + m)}}{\sigma}$$

(see also equation (11)).

The predictive ordinate of the data is given by

$$p_0^{\text{trunc}}(x_m) = p_0(x_m) \frac{\Phi\{z_m(\delta_L)\} - \Phi\{z_m(\delta_U)\}}{\Phi\{z_0(\delta_L)\} - \Phi\{z_0(\delta_U)\}}$$

where $p_0(x_m)$ is from expression (9). This predictive distribution does not appear to be integrable in closed form, and hence Box's test of prior–likelihood compatibility (Section 6.3) is not easily available, and also the chance of falling in critical regions cannot be explicitly calculated. Hence further predictive statements require numerical techniques for their computation.

A.3. *Mixture Priors*

We suppose that we have a prior that is a weighted mixture

$$p_0^{\text{mix}}(\delta) = w_0^A p_0^A(\delta) + w_0^B p_0^B(\delta)$$

where the components $p_0^A(\delta)$ and $p_0^B(\delta)$ may be standard, lump or truncated normal densities and $w_0^A + w_0^B = 1$. The posterior distribution is a mixture of the component posterior densities but with revised weights w_m^A and w_m^B, so that

$$p_m^{\text{mix}}(\delta) = w_m^A p_m^A(\delta) + w_m^B p_m^B(\delta)$$

where the weights are obtained from the odds version of Bayes theorem

$$\frac{w_m^A}{w_m^B} = \frac{p_0^A(x_m)}{p_0^B(x_m)} \frac{w_0^A}{w_0^B}$$

where $w_m^A + w_m^B = 1$, and the predictive ordinates $p_0^A(x_m)$ and $p_0^B(x_m)$ are obtained from expressions given previously. The distribution function of this posterior distribution is trivially obtained.

We also have the predictive ordinate

$$p_0^{\text{mix}}(x_m) = w_0^A p_0^A(x_m) + w_0^B p_0^B(x_m).$$

However, Box's procedure requires the distribution function of the predictive distribution, and hence is unavailable in closed form if at least one of the components is a truncated normal. After having observed data, predictive statements concerning future critical events may only be explicitly calculated if the analysis is not to include the prior, or there is no truncation in the prior.

APPENDIX B: INTERIM PREDICTIONS

By obtaining predictive distributions we can assess the chances that future data fall in critical regions such as A, B, D or E (Fig. 4). In our examples below, we consider the probability that the old treatment is clinically inferior (situation D), but it should be clear how to derive the formulae for other criteria of interest. We consider the fully Bayesian, mixed and likelihood approaches.

B.1. *Fully Bayesian Approach: Prior used in Predictions and Analysis*

For the lower extreme of the range of equivalence δ_I, we are interested in values of (X_n, x_m) that lead to $P_{m+n}(\delta_I) < \epsilon$, where P_{m+n} is the distribution function corresponding to the future posterior density $p_{m+n}(\delta) = p(\delta \mid X_n, x_m)$. For a simple normal prior we have that

$$p_{m+n}(\delta) = \phi\left(\frac{n_0\delta_0 + mx_m + nX_n}{n_0+m+n}, \frac{\sigma^2}{n_0+m+n}\right),$$

and so this critical event will occur if

$$X_n > \frac{n_0+m+n}{n}\delta_I - \frac{\sqrt{(n_0+m+n)}}{n} z_\epsilon \sigma - \frac{n_0\delta_0 + mx_m}{n}. \tag{10}$$

A similar expression for the critical value that leads to $P_{m+n}(\delta_S) > 1-\epsilon$ can be obtained.

Using the predictive distribution (6) we obtain that the predictive probability that the future posterior tail area $P_{m+n}(\delta_I)$ will be less than ϵ is

$$p_m\{P_{m+n}(\delta_I) < \epsilon\} = \Phi\left\{z_m(\delta_I)\sqrt{\left(\frac{n_0+m+n}{n}\right)} + z_\epsilon\sqrt{\left(\frac{n_0+m}{n}\right)}\right\} \tag{11}$$

where

$$z_m(\delta_I) = \left(\frac{n_0\delta_0 + mx_m}{n_0+m} - \delta_I\right)\frac{\sqrt{(n_0+m)}}{\sigma}$$

is the statistic from which the current tail area is derived.

This analysis also yields useful expressions for those unwilling to incorporate prior distributions into their predictions or analyses. If we take $n_0 = 0$, the improper reference prior that yields a posterior distribution that is identical with the normalized likelihood, we obtain from equation (6) a predictive distribution

$$p_m^{\text{ref}}(X_n) = \phi\left\{X_n \mid x_m, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right\}. \tag{12}$$

B.2. *Mixed Approach: Prior used for Predictions but not Analysis*

If we wish not to include prior opinion in reporting, then the final inference will be based on a reference posterior distribution denoted $p_{m+n}^{\text{ref}}(\delta)$, and hence future critical regions are derived by setting $n_0 = 0$ in inequality (10). Using the predictive distribution (6) gives predictive probability

$$p_m\{P_{m+n}^{\text{ref}}(\delta_I) < \epsilon\} = \Phi\left[z_0(\delta_I)\sqrt{\left\{\frac{n_0 n}{(n_0+m)(n_0+m+n)}\right\}}\right.$$

$$\left. + z_m^{\text{ref}}(\delta_I)\sqrt{\left\{\frac{m(n_0+m+n)}{n(n_0+m)}\right\}} + z_\epsilon\sqrt{\left\{\frac{(m+n)(n_0+m)}{n(n_0+m+n)}\right\}}\right] \tag{13}$$

where

$$z_m^{\mathrm{ref}}(\delta_1) = (x_m - \delta_1)\sqrt{m}/\sigma$$

is the current standardized statistic using the reference prior. A particular example of equation (13) occurs when we have no data so far ($m = 0$) from which we obtain expression (8) used in predictive power assessments. If a classical stochastic curtailment procedure is desired (although it is not recommended by us), predictions conditional on a particular value δ' are obtained by placing a lump prior on δ', in which case letting $n_0 \to \infty$ in equation (13) gives

$$\Phi\left\{(\delta' - \delta_1)\frac{\sqrt{n}}{\sigma} + z_m^{\mathrm{ref}}(\delta_1)\sqrt{\left(\frac{m}{n}\right)} + z_\epsilon\sqrt{\left(\frac{m+n}{n}\right)}\right\}. \tag{14}$$

B.3. *Likelihood Approach: Prior Ignored Both in Predictions and in Analysis*

Choi and Pepple (1989), Grieve (1991), Frei *et al.* (1987) and Hilsenbeck (1988) all discussed making predictions solely on the basis of the data so far. All effects of the prior can be removed by setting $n_0 = 0$ in either of expressions (13) or (11) to give

$$p_m^{\mathrm{ref}}\{P_{m+n}^{\mathrm{ref}}(\delta_1) < \epsilon\} = \Phi\left\{z_m^{\mathrm{ref}}(\delta_1)\sqrt{\left(\frac{m+n}{n}\right)} + z_\epsilon\sqrt{\left(\frac{m}{n}\right)}\right\} \tag{15}$$

which matches expression 4.6 of Jennison and Turnbull (1990). The interesting aspect of equation (15), noted in Spiegelhalter *et al.* (1993), is that it can be expressed solely in terms of the current standardized test statistic $z = z_m^{\mathrm{ref}}(\delta_1)$ and the fraction $f = m/(m+n)$ of the trial so far completed, to give the probability that the future tail area below δ_1 is less than ϵ:

$$p_m^{\mathrm{ref}}\{P_{m+n}^{\mathrm{ref}}(\delta_1) < \epsilon\} = \Phi\left\{\frac{z + z_\epsilon\sqrt{f}}{\sqrt{(1-f)}}\right\}. \tag{16}$$

This provides an exceptionally simple tool for those carrying out formal or informal interim analyses, who wish to make predictions based solely on the data so far. Often the consequences of such an analysis are depressing to investigators: if, say, they are half way through a study ($f = 0.5$), and their sample mean is currently one standard error from the null hypothesis ($z = 1$), they may feel that they are well on the way to success, whereas equation (16) will tell them that there is only a $\Phi(\sqrt{2} - 1.96) = 29\%$ chance that they will eventually be able to report that the final tail area is less than 0.05. Graphical figures showing the predictive probability of the eventual tail area falling below 0.025 for a range of current z-values and fractions of trial completed are given in Spiegelhalter *et al.* (1993).

Expression (16) can give rise to other useful predictive statements. Setting $\epsilon = \Phi^{-1}(-z)$ provides the predictive probability that the eventual tail area below δ_1 will be smaller than its current value $\Phi^{-1}(-z)$, whereas $\epsilon = 0.5$ gives the predictive probability that the sign of Z will stay the same. We see, for example, that for our investigator with $f = 0.5$ and $z = 1$ there is a 34% chance, based just on the data so far, that the tail area will be larger than its current value of 0.16, whereas there is an 8% chance that the final effect will point in the opposite direction. It is also straightforward to derive the full predictive distribution function for the future tail area by considering ϵ as a random variable in expression (15).

REFERENCES

AIMS Trial Study Group (1988) Effect of intravenous APSAC on mortality after acute myocardial infarction: preliminary report of a placebo-controlled clinical trial. *Lancet*, i, 545–549.

Armitage, P. A. (1985) The search for optimality in clinical trials. *Int. Statist. Rev.*, **53**, 15–24.
——(1989) Inference and decision in clinical trials. *J. Clin. Epidem.*, **42**, 293–299.
Becker, R. A., Chambers, J. M. and Wilks, A. R. (1990) *The New S Language*. Belmont: Wadsworth and Brooks/Cole.
Berry, D. A. (1987) Interim analysis in clinical trials: the role of the likelihood principle. *Am. Statistn*, **41**, 117–122.
BHAT Research Group (1981) Beta-blocker heart attack trial design features. *Contr. Clin. Trials*, **2**, 275–285.
Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A, **143**, 383–430.
Brown, B. W., Herson, J., Atkinson, N. and Rozell, M. E. (1987) Projection from previous studies: a Bayesian and frequentist compromise. *Contr. Clin. Trials*, **8**, 29–44.
Byar, D. P., Schoenfeld, D. A., Green, S. B., Amato, D. A., Davis, R., De Gruttola, V., Finkelstein, D. M., Gatsonis, C., Gelber, R. D., Lagakos, S., Lefkopoulou, M., Tsiatis, A. A., Zelen, M., Peto, J., Freedman, L. S., Gail, M., Simon, R., Ellenberg, S. S., Anderson, J. R., Collins, R., Peto, R. and Peto, T. (1990) Design considerations for AIDS trials. *New Engl. J. Med.*, **323**, 1343–1348.
Byar, D. P., Simon, R. M., Friedewald, W. T., Schlesselman, J. J., DeMets, D. L., Ellenberg, J. H., Gail, M. H. and Ware, J. H. (1976) Randomized clinical trials: perspective on some recent ideas. *New Engl. J. Med.*, **295**, 74–80.
Carlin, B. P., Chaloner, K., Church, T., Louis, T. A. and Matts, J. P. (1993) Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. *Statistician*, **42**, 355–367.
Carlin, J. B. (1992) Meta-analysis for 2×2 tables: a Bayesian approach. *Statist. Med.*, **11**, 141–158.
Chaloner, K., Church, T., Louis, T. A. and Matts, J. P. (1993) Graphical elicitation of a prior distribution for a clinical trial. *Statistician*, **42**, 341–353.
Choi, S. C. and Pepple, P. A. (1989) Monitoring clinical trials based on predictive probability of significance. *Biometrics*, **45**, 317–323.
Cornfield, J. (1966) Sequential trials, sequential analysis and the likelihood principle. *Am. Statistn*, **20**, 18–23.
Crook, J. F. and Good, I. J. (1982) The powers and strengths of tests for multinomials and contingency tables. *J. Am. Statist. Ass.*, **77**, 793–802.
Dixon, D. O. and Simon, R. (1992) Bayesian subset analysis in a colorectal cancer clinical trial. *Statist. Med.*, **11**, 13–22.
Dunnett, C. W. and Gent, M. (1977) Significance testing to establish equivalence between treatments with special reference to data in the form of 2×2 tables. *Biometrics*, **33**, 593–602.
Easterbrook, P. J., Berlin, J. A., Gopalan, R. and Matthews, D. R. (1991) Publication bias in clinical research. *Lancet*, **337**, 867–872.
Errington, R. D., Ashby, D., Gore, S. M., Abrams, K. R., Myint, S., Bonnett, D. E., Blake, S. W. and Saxton, T. E. (1991) High energy neutron treatment for pelvic cancers: study stopped because of increased mortality. *Br. Med. J.*, **302**, 1045–1051.
Fleiss, J. L. (1981) *Statistical Methods for Rates and Proportions*, 2nd edn. New York: Wiley.
Fleming, T. R. and Watelet, L. F. (1989) Approaches to monitoring clinical trials. *J. Natn. Cancer Inst.*, **81**, 188–193.
Freedman, L. S., Lowe, D. and Macaskill, P. (1984) Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, **40**, 575–586.
Freedman, L. S. and Spiegelhalter, D. J. (1983) The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Statistician*, **32**, 153–160.
——(1989) Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Contr. Clin. Trials*, **10**, 357–367.
——(1992) Application of Bayesian statistics to decision-making during a clinical trial. *Statist. Med.*, **11**, 23–36.
Frei, A., Cottier, C., Wunderlich, P. and Ludin, E. (1987) Glycerol and dextran combined in the therapy of acute stroke. *Stroke*, **18**, 373–379.
Genest, C. and Zidek, J. V. (1986) Combining probability distributions: a critique and an annotated bibliography. *Statist. Sci.*, **1**, 114–148.
Greenhouse, J. B. (1992) On some applications of Bayesian methods in cancer clinical trials. *Statist. Med.*, **11**, 37–54.
Grieve, A. P. (1991) Predictive probability in clinical trials. *Biometrics*, **47**, 323–330.

Grossman, J., Parmar, M. K. B., Spiegelhalter, D. J. and Freedman, L. S. (1994) Unified hypothesis testing, point estimation and interval estimation for group sequential clinical trials. *Statist. Med.*, **13**, in the press.

Halperin, M., Lan, K. K. G., Ware, J. H., Johnson, N. J. and DeMets, D. L. (1982) An aid to data monitoring in long-term clinical trials. *Contr. Clin. Trials*, **3**, 311–323.

Hilsenbeck, S. G. (1988) Early termination of a phase II clinical trial. *Contr. Clin. Trials*, **9**, 177–188.

Ingelfinger, J. A., Mosteller, F., Thibodeau, L. A. and Ware, J. H. (1987) *Biostatistics in Clinical Medicine*, 2nd edn. New York: Macmillan.

Jennison, C. and Turnbull, B. W. (1990) Statistical approaches to interim monitoring of medical trials: a review and commentary. *Statist. Sci.*, **5**, 299–317.

Kass, R. E. and Greenhouse, J. B. (1989) Comments on 'Investigating therapies of potentially great benefit: ECMO' (by J. H. Ware). *Statist. Sci.*, **4**, 310–317.

Lachin, J. M. (1981) Introduction to sample size determination and power analysis for clinical trials. *Contr. Clin. Trials*, **1**, 13–28.

Laurie, J. A., Moertel, C. G., Fleming, T. R., Wieand, H. S., Leigh, J. E., Rubin, J., McCormack, G., Gerstner, J. B., Krook, J. E., Mailliard, J., Twllo, D. I., Merton, R. F., Tschelter, L. K. and Barlow, J. F. (1989) Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. *J. Clin. Onc.*, **7**, 1447–1456.

Lee, P. M. (1987) *Bayesian Statistics: an Introduction*. London: Arnold.

Lindley, D. V. and Singpurwalla, N. D. (1991) On the evidence needed to reach agreed action between adversaries, with application to acceptance sampling. *J. Am. Statist. Ass.*, **86**, 933–937.

Makuch, R. W. and Simon, R. (1982) Sample size requirements for comparing time-to-failure among k treatment groups. *J. Chron. Dis.*, **35**, 861–867.

McPherson, C. K. (1974) Statistics: the problem of examining accumulating data more than once. *New Engl. J. Med.*, **290**, 501–502.

Mehta, C. R. and Cain, K. C. (1984) Charts for the early stopping of pilot studies. *J. Clin. Onc.*, **2**, 676–682.

Meier, P. (1975) Statistics and medical experimentation. *Biometrics*, **31**, 511–529.

Moertel, C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie, J. A., Goodman, P. J., Ungerleider, J. S., Emerson, W. A., Tormey, D. C., Glick, J. H., Veeder, M. H. and Mailliard, J. A. (1990) Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New Engl. J. Med.*, **322**, 352–358.

Moussa, M. A. A. (1989) Exact, conditional and predictive power in planning clinical trials. *Contr. Clin. Trials*, **10**, 378–385.

MRC Urological Working Party (1985) Intravesical thiotepa for superficial bladder tumors: an MRC randomized study. *Br. J. Urol.*, **57**, 680–689.

MRC Working Party on Misonidazole in Cancer of the Cervix (1983) The Medical Research Council Trial of misonidazole in carcinoma of the uterine cervix. *Br. J. Radiol.*, **57**, 491–499.

MRC Working Party on Misonidazole in Gliomas (1982) A study of the effect of misonidazole in conjunction with radiotherapy for the treatment of grades 3 and 4 astrocytomas. *Br. J. Radiol.*, **56**, 673–682.

MRC Working Party on Misonidazole in Head and Neck Cancer (1983) A study of the effect of misonidazole in conjunction with radiotherapy for the treatment of head and neck cancer. *Br. J. Radiol.*, **57**, 585–595.

O'Brien, P. C. and Fleming, T. R. (1979) A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556.

Parmar, M. K. B., Spiegelhalter, D. J. and Freedman, L. S. (1994) The CHART trials: design and monitoring. *Statist. Med.*, **13**, in the press.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I, Introduction and design. *Br. J. Cancer*, **34**, 585–612.

Pocock, S. J. (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191–199.

——(1983) *Clinical Trials: a Practical Approach*. Chichester: Wiley.

Pocock, S. J. and Hughes, M. J. (1989) Practical problems in interim analyses, with particular regard to estimation. *Contr. Clin. Trials*, **10**, 209S–221S.

Pocock, S. and Spiegelhalter, D. J. (1992) Grampian region early anistreplase trial. *Br. Med. J.*, **305**, 1015.

Poon, M. A., O'Connell, M. J., Moertel, C. G., Wieand, H. S., Cullin, S. A., Everson, L. K., Krook, J. E., Mailliard, J. A., Laurie, J. A., Tschelter, L. K. and Wiesenfeld, M. (1989) Biochemical modulation of fluorouracil: evidence of significant improvement of survival and quality of life in patients with advanced colorectal carcinoma. *J. Clin. Onc.*, **7**, 1407–1418.

Rosenbaum, P. R. and Rubin, D. (1984) Sensitivity of Bayes inference with data-dependent stopping rules. *Am. Statistn*, **38**, 106–109.

Sasahara, A. A., Cole, T. M., Ederer, F., Murray, J. A., Wenger, N. K., Sherry, S. and Stengle, J. M. (1973) Urokinase pulmonary embolism trial: a national cooperative study. *Circulation*, **47**, suppl. 2, 1–108.

Schwartz, D., Flamant, R. and Lellouch, J. (1980) *Clinical Trials* (translated by M. J. R. Healy). London: Academic Press.

Spiegelhalter, D. J. and Freedman, L. S. (1986) A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statist. Med.*, **5**, 1–13.

——(1988) Bayesian approaches to clinical trials. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 453–477. Oxford: Oxford University Press.

Spiegelhalter, D. J., Freedman, L. S. and Blackburn, P. R. (1986) Monitoring clinical trials: conditional or predictive power? *Contr. Clin. Trials*, **7**, 8–17.

Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. (1993) Applying Bayesian thinking in drug development and clinical trials. *Statist. Med.*, **12**, 1501–1511.

Ten Centre Study Group (1987) Ten centre study of artificial surfactant (artificial lung expanding compound) in very premature babies. *Br. Med. J.*, **294**, 991–996.

Tsiatis, A. A. (1981) The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*, **68**, 311–315.

United Kingdom Coordinating Committee on Cancer Research (1989) Axis protocol. *Technical Report*. Medical Research Council Cancer Trials Office, Cambridge.

Ware, J. H., Muller, J. E. and Braunwald, E. (1985) The futility index: an approach to the cost-effective termination of randomized clinical trials. *Am. J. Med.*, **78**, 635–643.

Whitehead, J. (1992) *The Design and Analysis of Sequential Clinical Trials*. Chichester: Horwood.

Yusuf, S., Collins, R. and Peto, R. (1984) Why do we need some large and simple randomized trials. *Statist. Med.*, **3**, 409–420.

Yusuf, S., Peto, R., Lewis, J., Collins, R. and Sleight, P. (1985) Beta-blockade during and after myocardial infarction: an overview of the randomized trials. *Prog. Cardvasc. Dis.*, **27**, 335–371.

DISCUSSION OF THE PAPER BY SPIEGELHALTER, FREEDMAN AND PARMAR

A. P. Grieve (ZENECA Pharmaceuticals, Macclesfield): I welcome the opportunity that this paper provides for a public debate of the appropriateness of the application of Bayesian ideas to clinical trials. It is timely on two counts. First, there is these days a far greater acceptance of the use of Bayesian methods in general. Second on the drug regulatory front there is now an explicit acceptance of Bayesian methodology. To illustrate:

> 'Agencies should also be receptive to the use of various newer statistical and pharmacometric techniques such as Bayesian and population methods, modelling, and pharmacokinetic–pharmacodynamic approaches' (International Conference on Harmonisation, 1993).

> 'Although this Note for Guidance is written largely from a classical (frequentist) viewpoint, the use of Bayesian or other well-argued approaches is quite acceptable' (European Committee for Proprietary Medicinal Products, 1993).

I shall concentrate my remarks on the choice, and use, of prior distributions and on the monitoring of studies as data accumulate.

Grieve (1994) points out that the statistician's job is not over when the analysis of a trial has been completed and a report written. At that stage thought must be given to the transmission of the information to diverse groups of remote clients (Hildreth, 1963). Examples of remote clients are experimenters, reviewers and individual doctors (Spiegelhalter and Freedman, 1988) all of whom will have different motivations and knowledge and therefore, not surprisingly, will interpret results differently. Hildreth proposes a number of packages of information which may be 'efficiently' transmitted to remote clients, one of which is the set of posterior distributions derived from a series of 'representative prior distributions',

the authors' community of priors. It may be necessary to provide more than one parcel of information and this is acknowledged by the authors in Section 8.4. I believe that this community should be as wide as possible. In particular when only subjective clinical opinion is available I would prefer to see the posterior from each of the individual clinical priors presented, as I find averaging of opinions difficult to accept as an 'estimated opinion of a typical participating clinician'. The priors in Table 1 suggest that the group mean is only typical of clinician 6 who is, in a sense, extreme in having the most diffuse prior. While on this example, surely sceptical and enthusiastic priors would be 'better' provided by clinicians 1 and 9 rather than the artificial constructs in Sections 4.1.3 and 4.1.4?

What is a prior for? One use, as the authors rightly point out, is to allow a 'formal basis for the use of external evidence'. To illustrate, Smithells and Sheppard (1980) in reply to a comment on a paper on vitamin supplementation to prevent neural tube defects, wrote:

> 'In our paper we subscribe to no belief, reach no conclusions, and offer four possible interpretations . . . . We are not trying to "convince the sceptics", amongst whom we count ourselves.'

In fact in the original paper Smithells *et al.* (1980) demolished three of the possible four interpretations on the basis of prior data and knowledge and this would seem an ideal case to merge prior beliefs with data to judge the relative merits of the four interpretations.

A second use is as means of shrinkage or as a yardstick against which a surprising finding may be measured (Pocock and Spiegelhalter, 1992). Browner and Newman (1986) used a somewhat similar argument to counter the move in medicine to an unthinking use of confidence intervals:

> 'Confidence intervals are entirely data-based . . . because they do not take into account the prior likelihood of the results, confidence intervals that are inconsistent with scientific knowledge cannot be viewed with confidence'.

Turning to monitoring, the concept of spending the type I error (Lan and DeMets, 1983) and the exhortation to regard stopping rules only as guidelines (DeMets, 1984) have always seemed to me to give rise to difficult issues. In most instances adjusting for the number of analyses undertaken has concentrated on the danger of too many analyses. However, if we regard stopping rules solely as guidelines then a monitoring committee may, for extremely good reasons, choose not to terminate a study despite a significant result. Have we spent the allocated proportion of the type I error for this analysis? Should we now use $K-1$ rather than K for the number of interim analyses? Had we used $K-1$ would previous results which were not significant now be significant? The authors support the idea of monitoring schemes which protect against early stopping, i.e. schemes in which the allocated type I error increases as the amount of data increases. I would be interested in knowing how they view Pitman's results, reported by Freeman (1993), which suggest that, rather than increasing, the allocated type I error should decrease.

The authors have continued a recent tradition of eschewing controversy in favour of an exposition highlighting the practical advantages of a Bayesian approach. I have always felt this to be more fruitful than an ideological slanging match, though the latter is often more entertaining. This view is not new. Over 300 years ago the following advice was given:

> 'And I think Mr. Bayes, when he hath had time to cool his thoughts, may be trusted yet with that consideration, and to compute whether the good that he hath done by Railing do countervaile the damage which both he in particular and the cause he labours, have suffered by it' (Andrew Marvell—*The Rehearsal Transpros'd*—1672).

It gives me great pleasure to propose a vote of thanks to the authors.

Stuart J. Pocock (London School of Hygiene and Tropical Medicine): I am delighted to have this opportunity to second the vote of thanks to the authors. We have received a splendid paper illustrating all the good things about the role of Bayesian methods in the design, monitoring, analysis and interpretation of clinical trials.

However, I think that it is appropriate for me to adopt a constructively critical stance, under the heading 'thoughts of a frequentist sceptic'. In reality, this is made somewhat difficult by the existence of my own Bayesian track record, albeit erratic. My first Bayesian article in 1976 was followed by a 12-year lapse into frequentism. A burst of Bayesian papers in the late 1980s, in collaboration with our unit's resident Bayesian, Michael Hughes, culminated in a letter on the Sellafield case–control study to the *British Medical Journal* (which was rejected!). A further lapse into frequentism might have been long term but for the stimulus of brief collaboration with one of today's authors, leading to another letter in the *British Medical Journal* (successful this time!).

So one might consider whether we traditional frequentists are prone to become Bayesians when faced with practical research questions, especially in clinical trials, which do not readily lend themselves to frequentist solutions. Indeed one might adopt the term 'closet Bayesian' for statisticians and other scientists who adopt strategies in study design and data interpretation which include concepts of prior belief, but who do not explicitly express them in a formal Bayesian framework.

Before concentrating on my more negative comments, may I express three specific positive views?

(a) Bayesian conceptualization is very useful in study design, and in particular the oversimplification of frequentist power calculations might usefully be enhanced by methods of sample size determination that incorporate prior belief. Perhaps the authors might have given more space to this topic.

(b) It is useful to be a temporary Bayesian when faced with a surprising frequentist result. For instance, the Grampian region early anistreplase trial compared home *versus* hospital thrombolysis after myocardial infarction and the mortality comparison of 13/163 *versus* 23/148 deaths respectively lead to an exceedingly promising 49% risk reduction ($p = 0.04$). Help, cries the knowledgeable cardiologist! Such a great benefit from giving the same treatment earlier simply does not make sense. In steps the Bayesian, in this case Pocock and Spiegelhalter (1992), to incorporate retrospectively a 'realistic prior' which pulls back the posterior distribution into more moderate territory. The trouble is that this is rather a Bayesian afterthought (creating the prior *after* the data have arrived which might be deemed 'cheating') but the principle is nevertheless a valuable quantitative reinforcement of common sense.

(c) It is useful that every medical statistics unit should have a resident Bayesian. Their philosophical stimulus is a valuable adjunct to the frequentist's armoury of techniques, which can appear cumbersome (control of type I error in multiple look-at data) and is often misinterpreted (how many non-statisticians really understand p-values and confidence intervals?).

However, I do see several practical and philosophical limitations in the Bayesian approach which need to be addressed.

(a) The specification of priors raises practical problems. We have heard a helpful account of how priors can be elicited (Section 4) but it is difficult to specify a meaningful prior that will be accepted by all interested parties. Incorporating clinical opinion takes a large effort and is of course heavily dependent on who is asked. The hardened sceptical trialist, the hopeful clinician and the optimistic pharmaceutical company will inevitably have grossly different priors. Averaging priors may cover up discrepancies, while handling multiple priors is cumbersome.

(b) The Bayesian approach lacks objectivity and reproducibility in that no two Bayesian analyses of the same data need agree because of different choices of prior. Hence, it is undesirable to have Bayesian methods replacing frequentist inference, since the objective rigour of the latter is vital in separating factual evidence from clinical opinion. An interesting quote from Fisher (1950) may be of relevance:

> 'We should recognise Bayes' greatness in perceiving the problem (of inference) to be solved, in making an ingenious attempt at its solutions, and finally in realising more clearly than many subsequent writers the underlying weakness of his attempt'.

I endorse the authors' plea to separate results in journal papers from Bayes interpretation (Section 8.4). However, I think the former still needs frequentist methods and, given that journal space is always at a premium, Bayesians need to enhance the art of concise presentation, e.g. graphs of prior, posterior and predictive distributions will often not be allowable.

(c) Both frequentist and Bayesian inference face difficulties as regards repeated looks at accumulating data. I do not wish to underrate the inadequacies of frequentist juggling of p-values. However, I do feel that Bayesian inference is flawed in this context in not facing the fact that inference and decision-making on any particular analysis should not be done in isolation but need to take account of the whole trial process, i.e. what previous and future analyses and related decisions were and will be intended. In the extreme, how can Bayesian inference over the course of a whole trial restrain the enthusiast who analyses his accumulating data daily in the continual search for a treatment difference? Although analogies between Bayesian and frequentist monitoring schemes have been made, Bayesian decision-making regrettably still lacks an equivalent to the frequentist control of type I error.

(d) Many statisticians working on clinical trials are very busy in coping with the design, analysis and interpretation of a wide range of studies. Hence, to function efficiently we need concise, well-understood non-controversial techniques for everyday statistical life. In practice, Bayesian methods take up too much time and space to be used routinely in clinical trials. Thus, I envisage Bayesian methods as primarily for 'special occasions', i.e. major trials in which detailed planning or interpretation can valuably incorporate prior belief or information as an aid to overall and personal decision-making.

(e) Given that it is inevitably difficult (if not impossible) to function as a total Bayesian in applied clinical trials, the real battleground is to persuade frequentists to incorporate Bayesian methods into their armoury of techniques. Thus, non-Bayesians need greater help in overcoming computational difficulties, in adapting Bayesian methods to more complex problems (e.g. covariate adjustment and repeated measures) and in learning how to communicate Bayesian findings to non-specialists. Although Bayesians can reply that software exists, complex problems are solvable and doctors think like Bayesians anyway, I feel that these practicalities are major deterrents to a more widespread use of Bayesian methods.

(f) One commendable aspect of this paper is that it brings Bayesian methodology closer to the real world of clinical trials. In the past, there have sometimes been expressions of a more naïve Bayesian philosophy that have not given sufficient concern to the value of randomization, the need to avoid bias and other essentials of clinical trials.

(g) There is always a possibility that my sceptical remarks reflect an underlying feeling of envy. Bayesians appear often to have more fun (and that is not fair, is it!). The very essence of Bayesian thinking is to adopt a personal attitude to data. Expressions of prior belief involve a more philosophical outlook which is perhaps more relaxing than the frequentist's worry about his overall type I error. I suspect also that most Bayesians are less worn down by data analysis, deadlines and doctors.

In conclusion, I would actively encourage a wider use of Bayesian methodology in clinical trials since only by such practical experience can its true potential be properly evaluated. This paper is a most valuable step in the right direction; I hope that it will provoke more trialists into experimenting with Bayes and I am pleased to second the vote of thanks.

The vote of thanks was passed by acclamation.

Keith Abrams (University of Leicester): I would like to congratulate the authors on a splendid review and demonstration of how Bayesian methods can be applied to clinical trials. I wish to raise two points.

The first regards the use of the predictive distribution in the stopping of a trial. Although I do not entirely agree with the authors in Section 6.4 when they state that the monitoring of a trial should be based on the posterior distribution rather than the predictive distribution, I would like to give an example of a trial in which the decision to stop was based on summary statistics from a predictive distribution. A recently stopped Cancer Research Campaign trial in colorectal cancer was designed to assess the role that carcino-embryonic antigen (CEA) directed treatment of tumour recurrence should play in routine clinical practice. *A priori* a treatment–policy difference yielding a log-hazard ratio of 0.42 was considered to be necessary before CEA-directed surgery would be used routinely. Fig. 6 shows the posterior predictive density, using a vague uniform prior distribution, for the log-hazard ratio, after having observed 171 events, whereas the trial had been planned to observe 425. The posterior predictive probability of observing a log-hazard ratio of greater than 0.42 was sufficiently small (0.002) that further entry into the trial was considered unethical and the trial was stopped.

The second point that I would like to raise is that of modelling failure time data. In many clinical trials, and especially in cancer, the time to a prespecified event is the main outcome measure. Although the 'normal theory' models outlined in this paper can accommodate such data, allowance for covariates is a problem. One solution to this problem is to use fully parametric proportional hazards models. Algebraically these models take the form

$$\lambda(t|\boldsymbol{\theta}, \boldsymbol{\beta}) = \lambda_0(t|\boldsymbol{\theta}) \exp(\boldsymbol{\beta}^{\mathrm{T}} X). \tag{17}$$

Estimation of base-line parameters $\boldsymbol{\theta}$ and regression parameters $\boldsymbol{\beta}$ in models such as equation (17) is relatively straightforward with Laplace approximations (Tierney and Kadane, 1986; Abrams *et al.*, 1994). These approximations require only the ability to optimize functions of several parameters and are relatively

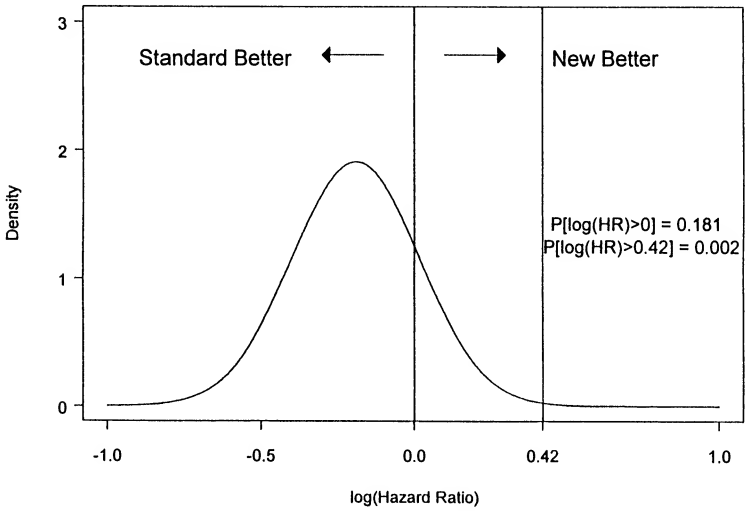

Fig. 6. Posterior predictive density for log-hazard ratio in the CRC CEA colorectal trial after observing $m = 171$ deaths out of an anticipated $n = 425$, using a vague uniform prior distribution (produced by kind permission of the CRC CEA Data Monitoring Committee)

straightforward to implement in environments such as S-PLUS (Statistical Sciences Inc., 1991), NAG (Numerical Algorithms Group, 1983) or LISP-STAT (Tierney, 1990).

Deborah Ashby (University of Liverpool): The authors are all respected both as statisticians and as trialists with many years of experience. This fortunate conjunction means that there are many areas where the work in this paper has had or is likely to have an effect. I shall restrict my comments to the regulation of medicines.

At the moment nearly all this work is conducted in a classical framework, and guidelines have tended to be couched in terms of classical procedures. One of the reservations about the use of Bayesian methods is the perceived subjective nature of the prior distributions, and so the thorough discussion in Section 4 of the role of prior evidence is especially welcome. However carefully a trial or set of trials has been done, the evidence must be judged against a background of other knowledge and experience. A graphical display of the main conclusions quantified as a function of prior beliefs is always welcome, but I particularly like the 'standardization' of the sceptical and enthusiastic priors, as useful yardsticks against which to judge the robustness of conclusions. However, these seem to depend for their specification on a 'classical' trial design, and I wonder what would the authors recommend in a more fully Bayesian setting?

There is now an emphasis on harmonization of regulatory procedures both within the European Community and beyond, in all areas, including statistical practice. I sense an increasing willingness among statisticians involved in drug regulation to explore the use of Bayesian approaches, and it is to be hoped that as guidelines are revised they explicitly include advice in this area. This paper makes an important contribution to the process, by setting out useful methods in a comprehensive but clear way, by providing easy-to-use software and most importantly by giving real examples that are models of good practice.

M. J. R. Healy (Institute of Education, London): This may be one of the most important papers presented to the Royal Statistical Society in recent years. For a long time, Bayesian methods have been widely discussed in the theoretical journals. Now they are beginning to reach the users of statistical methods and this paper makes them available in an important field of application.

That they are urgently needed can scarcely be questioned. The standard frequentist methodology, not only in clinical trials but across the range of medical statistics, must be regarded as a confidence trick. As statisticians we abandon to the user the responsibility for choosing the design parameters with the minimum of guidance. (Why the invariable 5% for α? Conditional on this, why the larger 10% or even 20% for β? Is it really more important not to make a fool of yourself than it is to discover something new?) We have nothing helpful to say about the confirmatory trial, even when P. M. Grundy showed many years ago that such a trial required quite special design considerations involving the

information initially available (Grundy *et al.* (1956); K. D. F. Tocher's contribution to the discussion anticipates much of today's paper).

The problem that I would like to pose to the authors, though, is serious. We have now been teaching frequentist techniques in medical schools and imposing them on journal papers for several decades. How do we confront the Kuhnian paradigm shift that the spread of subjectivist thinking represents? How do we go before the enormous user community and confess to them that we got it all wrong, that they have to relearn a discipline that they already find difficult and unfamiliar and that the heaven-sent computer packages that minimize the need for thought when analysing data must now be abandoned?

C. Jennison (University of Bath): I would like to draw attention to frequentist methodology for combining information on a study's major end point with other, possibly external, evidence. First, this was a key motivation for the development of repeated confidence intervals as a tool for interim monitoring (Jennison and Turnbull, 1989). Secondly, it is often desirable to combine information from two or more internal sources, e.g. treatment efficacy and safety, and Jennison and Turnbull (1993) have shown how to extend group sequential methods to monitoring such a bivariate response.

The authors' proposed stopping rule, based on the current 'credible interval', appears a rather arbitrary choice and I wonder whether the implicit question, 'Do the available data allow us to reach a conclusion with confidence $1-\epsilon$?', is the correct one to ask. I would rather ask 'Is the cost of continued sampling, financial or ethical, justified by the increased probability of reaching the correct conclusion?'. Utilities must be specified to answer this question and, although these are undoubtedly difficult to elicit, it is difficult to assess any stopping rule in their absence.

Although frequentist group sequential tests are often derived without explicit reference to utilities, good frequentist tests have deep connections to fully specified Bayes decision problems. Suppose that we specify the type I error, power at a given alternative and group sizes, and seek the group sequential test with minimum expected sample size, integrated over some parameter distribution. Complete class theorems (Brown *et al.*, 1980) imply that an optimal test must be the Bayes test for some combination of prior and utility function which ascribes costs to sampling and to making a wrong decision. Eales and Jennison (1992) exploited this relationship in computing optimal frequentist group sequential tests; they also suggest that experimenters should examine the decision problem solved by a test to see whether the implicit utilities are reasonable.

Whatever their philosophical status, repeated sampling properties are useful in calibrating stopping rules. There is a limited connection between posterior confidence levels and frequentist error probabilities, especially in the sequential setting. Thus, I am pleased that the authors consider frequentist properties of their methods. I would invite them to go a step further in

(a) showing that their stopping rule is at least an approximate solution to a Bayes decision problem and
(b) presenting the utility function implicit in the underlying decision problem as a means of assessing the suitability of their stopping rule.

J. A. Lewis (University of Kent, Canterbury): This paper is very welcome because it adopts an integrative and not a divisive approach to Bayesian and frequentist methods. It also places its material in a practical context. Most applied statisticians have little interest in confrontation between rival philosophies but have a keen interest in pragmatic solutions to real problems and hence will applaud the realistic attitude of the authors.

Of particular importance is Section 8.4. Here the authors propose separation of the results of the trial from their interpretation. This separation constitutes a major step forwards in increasing the potential acceptability of Bayesian methods. In the regulatory context, for example, where different interpretations of clinical trial results can critically affect access to enormous medical markets, it is vital to keep one's feet on the solid ground of data and facts. Robust and verifiable assertions which can be made entirely, or almost entirely, on the basis of the data constitute the area of potential full agreement. This area must be crystal clear and explicit. Only in this way can areas of disagreement be isolated and understood. The subsequent use of Bayesian methods to introduce external opinion and evidence may then help to resolve such disagreements.

We are shown in two examples how prior distributions can be based on results from earlier randomized trials (Section 3.2.3, Medical Research Council neutron therapy trial, and Section 4.2.1, beta-blocker heart attack trial). It is instructive to observe that in each of these cases the estimate of the treatment effect from the subsequent trial agreed very closely with the prior estimate from the earlier trials.

In the Medical Research Council trial prior clinical opinion was at much greater variance with the subsequent trial results. I am concerned that all too often insufficient attention is paid to identifying and summarizing earlier relevant data. Statisticians may be tempted to leave this task to clinical colleagues. In my experience a prior distribution based on a statistician's thorough quantitative assessment of the literature will nearly always have the edge over a clinician's judgment. This has the consequence that I am reluctant to attach any further mystique to 'clinical opinion' in the context of clinical trial work. I do not wish to encourage my medical colleagues to think that they have a direct line to the truth which is independent of previous data. I want their views to reflect previous data, but I fear that this may be an unrealistic wish.

D. V. Lindley (Minehead): The authors are to be congratulated on a splendid paper. Many Bayesians, myself included, had thought that, once the methodology was in place, the applications to an applied field, like clinical trials, would be straightforward. This is not so; there are real difficulties that have been successfully overcome in this paper. There is one point where I disagree.

The authors dismiss the use of expected utility as being unrealistic. In a sense this is right, for to have given a treatment of the decision aspect of clinical trials, in addition to the inference side, would have unreasonably lengthened the paper and demanded much research. Nevertheless, it must be recognized that clinical trials are not there for inference but to reach a decision, and the omission of their *raison d'être* is serious. In the long term, expected utility is realistic and, indeed, necessary.

It must be 25 years ago since I gave a seminar on sequential methods, including clinical trials, and was told, in the discussion, that the use of probability to express one's beliefs was unrealistic. The authors have demonstrated that, on the contrary, it is realistic. They are wrong about expected utility just as my discussants were about probability.

We should use Bayesian concepts firstly because they work and secondly because only with them is it possible to be coherent. To appreciate this, contemplate the situation when you have just been given some data. Why not just look at them and express an opinion about the parameter? Why go through elaborate calculations, whether frequentist or Bayesian? If we can assess a prior directly, why not a posterior? The answer is: to achieve coherence, to make all our beliefs cohere; in particular, to make our final beliefs cohere with those that went into our likelihood. This is why frequentist p-values, interpreted as beliefs in the null hypothesis, are unsound; they do not cohere.

The same reasoning applies to decisions as to beliefs. It is only by using expected utility that we can be sure that our actions will fit together sensibly. I suspect that the procedure, suggested in Fig. 4, of continuing with the trial until a tail area probability in the posterior is small, is just as incoherent as a belief based on the tail area, p-value. Or, if it is coherent, it implies an inept utility, such as one taking only values 0 and 1.

I hope that it will not take 25 years before the Royal Statistical Society has a paper, perhaps from a student of Spiegelhalter's, doing for utility what this does so well for probability.

David Machin (Medical Research Council Cancer Trials Office, Cambridge): The authors have described a persuasive case for the use of Bayesian methods in clinical trials. It is clear that such an approach can provide sensible procedures for assessing trials that need to be stopped early or extended, incorporate external information into the final assessment and cope with unscheduled analysis.

However, it is important to remember that, at the design stage of any trial, rather arbitrary decisions are made with respect to test size, power and the method, and frequency, of any interim analysis. A vital ethical requirement, at this stage, is that there must be considerable uncertainty about the size of δ (Machin, 1992). This uncertainty implies that, if the trial is to be conducted at all, any initial prior remains 'very vague' and is centred close to $\delta = 0$.

The prior distribution obtained from investigators tends to be centred around their optimistic belief rather than a summary of their knowledge of the current evidence (Freedman and Spiegelhalter, 1983; Freedman *et al.*, 1984). At best it will be a compound of anecdotal information, the published literature and personal experience. In certain circumstances such a prior may be obtained from a formal overview (Stewart and Parmar, 1993). One question is whether the point estimator from an overview is used as the prior or to supplement the investigator's prior.

As a trial progresses the external information available will be evolving and, in using a Bayesian approach, it is essential to stipulate how it is to be selected. One possibility is to confine attention to randomized clinical trials that are included on an established register (Fayers and Armitage, 1993).

There is a clear need for case-studies (Fayers *et al.*, 1994) and in a previous paper (Spiegelhalter *et al.* (1993), Fig. 2) the authors reported an investigator prior for a trial of the European Osteosarcoma Intergroup. Recruitment of patients to this trial was completed in 1993 with 407 patients, and the analysis is in preparation. It will provide a valuable first case-study of the methods discussed in this paper.

G. B. Newman (Abbots Langley): I would like to take advantage of the authors' consideration of alternative prior distributions to introduce the Bayesian bigot. The Bayesian bigot is a Bayesian who, whatever the outcome of the trial, does not believe that the treatment being tested is of any use. The need for a Bayesian bigot came when considering responses to a trial in test-tubes of the immune reaction to homeopathic treatments. These treatments are diluted beyond the point when we would expect a single active molecule to be present in the preparation. The experiments appear to show that a reaction is still present and this would leave a vast hole in current scientific theory and medical treatment (Davenas *et al.*, 1988). So great was the disbelief, or possibly the consequences, the journal *Nature* sent a team to investigate the experiment. The experiment has been repeated many times, most recently at University College London (Hirst *et al.*, 1994), where the data were analysed by J. Burridge using frequentist methods. It is clear that any Bayesian could have dismissed the original conclusion by placing a sufficiently small prior in the region of the apparent results. How unrealistic this bigoted Bayesian would have to be before the experiment, still to prefer his own beliefs after the experiment, is a useful measure. The inclusion in *Nature*'s investigating team of a magician and expert in scientific fraud suggests that our authors should include some prior probability that the evidence is not related to the hypothesis at all. My Bayesian bigot could then use this to maintain his beliefs.

J. A. Bather (University of Sussex, Brighton): I was pleased to find that the authors accept the principle of randomization in clinical trials, although mildly disappointed that they did not include a Bayesian justification. This might be based on extending prior distributions to cover subject variability and possible allocation bias but, if such a justification is possible, I suspect that it would be complicated and unconvincing.

It is surprising that the control of type I errors is such an important feature in Section 8 of the paper. The authors do not make clear whether they would otherwise wish to specify the stopping rule in advance. The choice of sample size or the imposition of a sequential stopping rule is part of the design of an experiment. For this reason it is important to state the intended rule for termination, in spite of possible intervention by the monitoring committee during the experiment. If the stopping rule is not defined, it is difficult to compare the design with other possible ways of carrying out a controlled experiment. For example, one cannot evaluate operating characteristics such as error probabilities and expected sample sizes in advance.

R. A. Bailey (Goldsmiths' College, London): I congratulate the authors on their clear, interesting paper with its pragmatic approach. I particularly like the idea of sceptical and enthusiastic priors, with the corollary that the trial must collect enough data, of sufficiently high quality, to convince people who start from different assumptions.

However, it does seem that, as so often in pro-Bayesian papers, what is being criticized is not the frequentist approach *per se* but inappropriate use of frequentist methodology. If neither clinicians nor the regulatory authorities believe the null hypothesis then it is an inappropriate null hypothesis. Is there any essential difference between using Bayesian methods with sceptical and enthusiastic priors and using frequentist methods looking for enough evidence to convince two people with different null hypotheses?

My only major worry about this paper is the contradiction between the often-stated need for, and belief in, randomization in the conduct of clinical trials and the absence of any mention of randomization in the more mathematical parts of the paper. The paper is about 'randomized trials', and the authors 'fully accept the need for randomization'. But why is randomization used? It does not appear to affect either the statistical modelling or the data analysis.

Call the two treatments A and B. Two possible schemes for allocating treatments are the systematic alternating schemes ABABABA ... and BABABAB .... Among the many commonly suggested methods of randomizing the treatment allocation are: random choice between the two systematic alternating schemes; tossing a coin for each patient; randomized blocks of, say, 10 patients; randomization to balance over covariates. Do the authors recommend their methods if one of the systematic allocation schemes is used? Do they recommend exactly the same analysis no matter which randomization method is used?

S. J. W. Evans (London Hospital Medical College): Many Bayesians seem to me to be rather like unmarried marriage guidance counsellors! To change the metaphor, they do not dirty their hands with real data. Spiegelhalter, Freedman and Parmar are among those who have helped with real data, and they are to be thoroughly congratulated.

Our real problem is how to change the paradigm, as Professor Healy has suggested. I wonder whether some incremental approaches can be tried.

The first approach is an elicitation of priors. Whatever the paradigm, this seems worthwhile in itself and should be encouraged, whatever we are going to do, whether frequentist or Bayesian. It is useful for prior simulation of trials as well as for sample size calculations.

In relation to this, as some people have said, it is important to distinguish between belief and belief supported by evidence. The beliefs that the clinicians have are very often entirely unsupported.

Secondly, it seems to me that meta-analysis priors should be a requirement for part of the introduction to papers. I would ask whether the meta-analysis itself should be done in a Bayesian method. Publication bias is likely to decrease the variance of the prior as well as to bias its direction when we look at published papers.

The third area where I think that the paradigm can be changed is that in many senses the Bayesian approach answers the problem that the scientist has: it 'scratches where the scientist itches'. The mathematician's approach, which has been that of the frequentist, has not been an answer to the problem for the scientist about what is happening in the real world. It is always slightly sad that the Bayesians have concentrated too much on the mathematics in some instances and not enough on the science.

Lastly, the authors of this paper are to be congratulated if we believe in the union of statisticians, in that it will be much more difficult for non-statisticians to carry out the analyses.

Niels Keiding (University of Copenhagen): It is often said that the possible introduction of Bayesian ideas into clinical trials (and biostatistics in general) will require an inordinate amount of reteaching. However, we should not forget that the basic frequentist concepts are almost impossible to accept by the uninitiated lay user of statistics. As soon as we statistical priests are out of the room, they say 'P really means the probability that the treatment has no effect', even if they have understood that we shall condemn them when we hear their sins. It is the convoluted frequentist concepts that are so difficult to teach in the first place.

The authors phrase their discussion of clinical trials as if these form part of science, leaving room for imagination and intelligent reasoning in their design and interpretation. However, many (most?) clinical trials are there to protect us in our capacity as current or future consumers of pharmaceutical drugs. We have appointed the regulating agencies to rein in claims of efficacy founded only on profit for the industry. In their often unrewarding efforts to complete this task for us, they need to prescribe the ritual dances that at least in the frequentist guise look so ridiculous. Here imagination is forbidden, so this cannot be science. Do the authors believe that the Bayesian approach allows less ridiculous guidelines from the regulating agencies?

The decisive point to me is that we should give our subject a chance to come closer to the concepts of the well-informed lay public. Let us see how the Bayesian approach works and keep our bets that regulators will still need to set arbitrary standards.

J. N. S. Matthews (University of Newcastle upon Tyne): We have heard much about priors and beliefs this evening, so I shall take up this ecclesiastical theme. In the 1970s, the Archbishop of Canterbury was asked to comment on two musicals then in the West End: Godspell and Jesus Christ Superstar. The Archbishop said that the former converted people *because* of itself, and Jesus Christ Superstar *in spite* of itself. Dr Ramsay seemed to think that this was a point in favour of Godspell. However, it has always struck me that it is actually much more convincing for a conversion to have happened via Jesus Christ Superstar. That seemed to be a much more important feature.

This evening's lesson is essentially along those lines, in that, if we can show that treatments can be changed on the basis of a sceptical prior, clearly this will be more convincing. Therefore—as we have already heard from numerous discussants—the role of the sceptical prior is very important.

My worry about the elicitation of these priors is that medical statisticians tend to become involved in particular projects essentially by invitation and accretion to certain groups of medical colleagues. The medical research world seems to me to be a much more combative and brutal world than our gentler environment. Consequently it seems that the medical statistician may not be all that well placed to elicit sceptical priors, as there may be many clinical opinions of which he is unaware. Therefore, in eliciting

these priors, strict attention must be paid to the principles of statistics concerning sampling, so that such bias may be avoided.

Sheila M. Gore (Medical Research Council Biostatistics Unit, Cambridge): Mr President, ladies and gentlemen, the authors have ably illustrated the need for greater formalism in the use of external evidence in the design, monitoring and reporting of randomized clinical trials. My questions mainly concern the range of equivalence, which is crucial for clinical practice but 'is not straightforward' to specify and 'will often change as a trial progresses'. No uncertainty over the range of equivalence is admitted: why, I wonder, have the authors eschewed a prior distribution on the range of equivalence?

I suggest a need to consider reference limits for the range of equivalence. These can be appealed to if, for example, treatment protagonists are minded to shift the goalposts (the trial's range of equivalence). An argument for suspension of randomization in the Medical Research Council's neutron therapy trial was that, even if the upper limit of the range of equivalence had been as low as 0.15 (as with aspirin prophylaxis to reduce myocardial infarction), there was little likelihood of high energy neutrons being preferable to photons, and therefore randomization should cease. The Clatterbridge Review Committee, uncomfortable with decision-making based on such confidence intervals, might have understood the authors' graphical formalism.

The range of equivalence and prior belief concerning an experimental treatment are used to justify randomization. I raise a concern about continued 1:1 randomization, apparently with informed patient consent, when the interim posterior distribution shifts in relation to the range of equivalence from [25, 50, 25] at the start of the trial to say [50, 30, 20] favouring the experimental treatment at interim analysis. A consumer principle of randomization (Gore, 1994) would allow patients and their doctors to select between limited randomization ratios, such as 2:1 or 1:1 or 1:2, the chosen randomization stratum being, of course, a patient covariate. As the trial progresses, and as new evidence becomes available, the proportion of patients choosing a particular randomization stratum may shift, at patients' or their doctors' instigation.

Stephen Senn (Ciba Pharmaceuticals Division, Basle): I welcome this clear and useful account of Bayesian approaches to planning and analysing clinical trials. I have the following criticisms.

The first example raises some questions regarding priors. The experts were 'knowledgeable about neutron therapy' but must either have been in ignorance of the results of the individual trials already run or incoherent or confused during elicitation. Any of these calls the utility of this 'knowledge' into question. It does not appear possible to take the data from the previous trials and to subtract them from the prior by using the inverse of equation (3) in any way which makes sense. Yet, by doing this, we ought to be able to reach some sort of palaeo-prior representing beliefs before trials were first run. Is this 'group clinical opinion' not an example of information which is worse than useless? If Bayesian updating is to become a regular feature of analysing clinical trials and if expert opinion is to become part of the process then, to avoid counting results twice, it will be vital to know who knew what when. Further, in Section 4.2.3 the suggestion is that, where there are no similar previous studies, subjective priors are needed: why? In Section 3.2.3 the posterior of a meta-analysis was entertained as a possible prior to the trial. But, whatever expectations went into planning the trials comprising the meta-analysis, they do not appear to be regarded as relevant to the calculation of *its* posterior which in turn becomes adequate as a prior. The practical lesson for me is that there are conjectures used in planning trials which are not usefully combined with their result when reporting them—a position which is natural for all frequentists.

Second, I am not entirely happy with all the claims made regarding the Bayesian approach's advantages, in particular, as regards 'equivalence'. Methods for handling this exist in frequentist formulations (Mehring, 1993) but references on equivalence trials (Makuch and Johnson, 1989) justify the oxymoron 'equivalence is different'. For example, failed experiments are more likely to produce apparent equivalence and if a Bayesian approach is to meet the claim of absorbing equivalence and superiority in a single coherent framework, in the way in which the summary claims, it needs to model the competence of experiments (Senn, 1993). When this is done, certain frequentist concerns are seen to be reasonable.

Jay Herson (Applied Logic Associates, Houston): The authors are to be congratulated on presenting such a comprehensive and persuasive paper on the use of Bayesian methods in clinical trials. I have long believed that frequentist methods are useful for 'pre-trial betting' (i.e. design of a clinical trial), whereas post-trial or intra-trial inference should always be based on likelihoods. The Bayesian framework

which allows for incorporation of prior information into likelihoods appears to be ideal. I also champion the use of frequentist principles to select between several Bayesian schemes and was pleased to see the authors apply that technique in Section 8.1.

Still, the greatest challenge for Bayesian application lies in the specification of the prior distribution. Publication bias affects the accuracy of both literature-based priors and expert-elicited priors. Most trials have multiple efficacy end points, but the priors that we choose for them may arise from a variety of populations of patients. How can the covariance structure be specified to relate to the population being studied?

I was surprised to see the authors suggest any type of averaging of priors in the CHART study when those resulting from the elicitation effort differed so widely. Are we not faced with a problem akin to finding a treatment × centre interaction here? The middle ground may not be representative of any expert's prior knowledge.

The 'community of priors' approach would be impractical for use by pharmaceutical industry statisticians. They would fear that regulators might choose that prior from their community that is least favourable to their drug as the only reasonable prior to use. A single defendable prior would be required by these researchers.

The next 10 years will see a rapid development and proliferation of Bayesian methods and software. Bayes methods will someday help researchers to decide not only when to terminate a trial but also when to begin a trial—perhaps by estimating the cost of the trial/unit decrease in posterior regret that would result. Indeed, Bayesian methods may someday help patients to make decisions on which type of treatment to choose among several alternatives available (e.g. breast cancer and acquired immune deficiency syndrome).

We have come a long way since Bayesian methods in clinical trials were first suggested in the early 1960s. In bridging theory and practice, we have certainly not reached the end, not even the beginning of the end but, perhaps, the end of the beginning.

Bruce W. Turnbull (Cornell University, Ithaca): Certainly there is a place for both 'pragmatic' Bayesian and frequentist approaches in the deliberations of a data monitoring committee (DMC). My own views concerning the Bayesian approach have been summarized in Jennison and Turnbull (1990). Although rigid stopping rules are not realistic nor recommended, there is something to say for having rigid stopping guidelines. In an ideal protocol, the end points of interest are specified in advance together with a grouping of them into categories—major, minor, etc. Stopping guidelines are specified for each category. The meeting of a criterion forces a formal discussion and vote by the DMC. At the discussion, quantitative information about δ, in the form of a repeated confidence interval, say, is considered along with other information obtained from outside the trial or from inside the trial but independent of the estimate of δ. Any more formal method whereby a 'community of priors' is updated according to this accumulating external information may not be practical. Also it could be quite confusing when reported to a scientific or public audience.

There are dangers in procedures like those described in Section 5.1. Ignoring the multiplicity leads to procedures with poor frequentist properties and this should be of concern, as the authors seem to agree in Section 8. The authors disdain a Bayesian decision theoretic approach. Apparently this is for two reasons:

(a) a realistic assessment of utilities is impossible (Section 3.1) and
(b) the consequences of continuing are of secondary importance (Section 6.4).

The counter-argument to (a) is that precise knowledge of utilities is not required to obtain the general *form* of the optimal procedure. Typically, the general form of the optimal procedure is like a frequentist group sequential procedure because it will account for the consequences of continuing (Eales and Jennison, 1992). Against (b), we assume that the consequences of continuing are of crucial importance at the start of the trial—so why not a little later into the trial?

There are other pressing statistical problems in clinical trials where a Bayesian approach could shed valuable insight. One concerns the multiplicity of end points; in one current trial there are eight efficacy end points and over 30 major safety end points! The frequentist approach strives to protect type I error and is unduly conservative. Another high priority statistical problem, especially common in disease prevention trials, is how to account for contamination, non-compliers and drop-outs, where an intent-to-treat analysis may be inappropriate or irrelevant.

The following contributions were received in writing after the meeting.

Douglas G. Altman (Imperial Cancer Research Fund, London): The authors (and many discussants) have demonstrated that the dichotomy between 'Bayesians' and 'frequentists' is increasingly inappropriate. The authors have made valuable suggestions regarding the analysis and interpretation of clinical trials, incorporating ideas from both schools of thought. I look forward to seeing their ideas on the presentation of results put into practice in medical journals. One of the main concerns about the Bayesian method is the specification of the prior distribution. Although this paper includes examples of how prior distributions can be obtained in various ways, I do not think that all worries will be alleviated. A key issue that is not really addressed is the reliability of the prior information. For example, the prior distributions of nine clinicians for the treatment benefit in the CHART trial (Table 1) gave a very similar distribution to that of clinician 6 alone (whose distribution had an unusually high spread). I am uncomfortable about the idea that an analysis would take equal account of one subjective opinion as of the combination of nine opinions. (There is also the design issue of whose opinions are sought, and how many.) Further, treating opinions as equally informative as the results of prior trials does not seem right either, nor does ignoring the question of whether the prior trials are strictly relevant, as in the CHART study. It seems highly desirable to take account of the source and reliability of the prior information, although it is by no means obvious how these aspects can be quantified.

If these methods become more common I wonder about the possibility that once clinicians understand how their opinions will affect the results of the trial they might feel that it was in their interest to be (even more) overoptimistic about the effectiveness of a new treatment than they are at present.

P. Armitage (Wallingford): The authors have given us many illuminating papers on this subject, of which the present is perhaps the most penetrating. I am impressed by the breadth of its coverage, the persuasive yet balanced style of its presentation and the authors' refusal to lose sight of practical constraints.

They favour the use of uninformative priors, or equivalently likelihood, for the public transmission of trial results, and this seems sensible. As they remark, the conclusions will usually be effectively the same as with an appropriate frequentist analysis. Among the consumers who might want to use their own priors to interpret trial results are public bodies such as regulatory authorities. The authors suggest that in judging efficacy results these authorities might use sceptical priors. I am not so sure. A firm would have cause for complaint if its apparently effective drug were downgraded because of general scepticism. Conversely, toxic effects should be taken at their face value, rather than minimized because they were not seen in other submissions.

A different problem is posed by another important public body, the United States National Institutes of Health (NIH). A 1993 bill requires that NIH-sponsored trials should provide evidence that research results apply equally to both sexes and to different minority groups. This would, of course, require impossibly large trials, and there is much anxiety about the implications of this legislation. Perhaps the NIH should announce that the results of a trial will be interpreted with a multivariate prior assuming high correlation between treatment effects for different strata. This would be equivalent to the current tacit assumption that interactions are ignorable without strong evidence to the contrary and would imply that results from one trial could reasonably be generalized to other populations.

A final point concerns early termination based on the various sorts of interval estimate shown in Fig. 4. Clearly, A and E would require termination on ethical grounds. The authors suggest that B and D may also justify termination, and this apparently happened in examples 1 and 4 (but not example 3). These situations present no ethical problem, because it remains unclear whether one treatment is superior or whether the parameter is in the equivalence zone. It would often be important to reduce this ambiguity by continuation of the trial: why stop prematurely?

A. C. Atkinson (London School of Economics and Political Science): The design of clinical trials was more mentioned in the presentation than it is in the published version of this stimulating paper. The design aspects of clinical trials do, however, raise interesting problems in the application of Bayesian methods.

If the responses are normally distributed an appropriate linear model is

$$E(Y) = X\beta + Z\gamma$$

where X is the design matrix representing treatment allocation, functions of β are of interest and Z is a matrix of prognostic factors. The simplest trial is that in which patients are randomized to two treatments and the contrast $\beta_1 - \beta_2$ is the parameter of importance. Many reported trials are of this form. However, it is a principle of the design of experiments (Fisher (1960), section 40) that nothing is lost in a factorial experiment by adding extra factors. A first question is to what extent is the addition of subsidiary factors standard practice?

With two treatments, randomization should provide satisfactory balance over Z. But, with a more complicated treatment structure, straightforward randomization may need an exorbitant number of trials to guarantee a reasonable degree of balance. It then seems natural to incorporate Z into the allocation. Atkinson (1982) described an extension to the biased coin designs of Efron (1971) which uses a randomized version of D_A-optimality for contrasts in β to provide designs which have a high probability of being reasonably balanced whenever the trial is stopped. Some methodological references to other non-adaptive allocation rules are given in section 22.4 of Atkinson and Donev (1992). Are such designs of practical importance?

There is nothing Bayesian in this. However, the response will often have a non-normal distribution, e.g. binomial. Then the optimum design will depend on the values of β and γ. The locally optimum designs of Box and Lucas (1959) for non-linear normal theory models correspond to the use of a point prior. The Bayesian extension is to designs in which the expectation of the design criterion is maximized over the prior distribution of the parameters. Fortunately the general equivalence theorem of optimum experimental design theory goes through so that standard methods of design construction are available. Examples for binary data are given by Chaloner and Larntz (1989) and in chapter 19 of Atkinson and Donev (1992). Is there any practical interest in such Bayesian designs for clinical trials? Is it diminished by the remarks in Section 2.1 which return the problem to the normal theory linear model and away from the need for the specification of priors?

Donald A. Berry (Duke University, Durham): This is another fine paper by this team of authors. I congratulate them for their success in applying Bayesian attitudes to the design and analysis of clinical trials. I especially like the way that they handle elicitation. I hope and believe that their efforts will help to bring these attitudes to the mainstream of clinical trials, with enormous benefit for all. The updating process is scientifically pleasing and represents a giant step in methodology. I have a single comment and it concerns the future direction.

I disagree that a decision theoretic approach using utilities is unrealistic in clinical trials: 'the consequences of any particular course of action are so uncertain that they make the meaningful specification of utilities rather speculative'. Speculation and assessing uncertainty are the stuff of the Bayesian approach. In deciding whether to stop a trial the authors spurn utilities and so 'are left with the conventional bench-marks such as 2.5% and 5%'. In my view, deciding whether to stop a trial requires considering why we are running it in the first place, and this means assessing utilities.

We conduct clinical trials to learn about the relative safety and efficacy of the therapies involved. Learning is important. But designing a trial for the sake of learning alone is not consistent with delivering good medicine. Information has costs. Trial participants may receive inferior therapy. Patients outside the trial may be ill served if the trial lasts too long since access to information from the trial would be delayed. In contrast, a trial that is stopped too soon may lead to the wrong conclusion and patients may be treated inappropriately as a result. Or, the trial may be stopped with results that are convincing to the investigators but not to practitioners or regulatory officials. Or, the discovery of and experimentation with new and innovative therapies may be inappropriately suspended. In the Bayesian approach each of these possibilities can be considered explicitly.

An example in which the various uncertainties are explicitly considered is carried out in Berry *et al.* (1992, 1994). The goal is to prevent cases of *haemophilus* influenzae type b among native American infants over the next N years. Explicitly considering such a horizon is an idea due to Anscombe (1963) and Colton (1963). The trial is stopped when continuing it (optimally) results in a greater number of expected cases.

Spiegelhalter and co-workers use an indifference region and 'conventional bench-marks' as surrogates for utilities. I encourage them to develop a fully Bayesian approach and I am sure that it will be as elegant as the development embodied in the current paper.

Marc E. Buyse (International Institute for Drug Development, Brussels, and Limburgs Universitair Centrum, Diepenbeek): The basic tenet of Bayesian approaches is that there may be something to gain

from 'mixing' data with some prior distribution. Let us examine how such a prior might reasonably be obtained.

(a) The prior might be elicited from expert clinicians. The example of neutron therapy chosen by Spiegelhalter and colleagues is somewhat chilling, since the prior distribution based on clinical opinion bears no resemblance whatsoever either to the results of a meta-analysis of similar trials or to the results of the new trial itself! Although it may be instructive to document such a phenomenal discrepancy, it may not be sensible to mix the trial data with a prior which is so blatantly irrelevant.
(b) The prior might be based on the results of a meta-analysis. As Stein suggested long ago (from an empirical Bayesian viewpoint), it may be sensible to shrink the extreme result observed in a single trial towards the mean of all results observed thus far. Even when a meta-analysis is available, however, *extrapolation* of its results should still be viewed with scepticism. A striking example occurred when all investigators involved in the overview of early breast cancer trials were asked to predict a plausible range for the 10-year survival benefits of adjuvant therapy based on the 5-year survival benefits. The 10-year benefits actually observed were far larger than even the most optimistic estimate (Early Breast Cancer Trialists' Collaborative Group, 1992)!
(c) The prior might reflect plain scepticism (with much of the prior mass on zero treatment effect), which implies that only extreme trial results will be deemed convincing. This is the view taken by regulatory authorities who require that two independent trials be significant at the 0.05 level (an overall significance level of 0.0025) before they grant market authorization to a new drug. We can wonder whether the heavy artillery of Bayesianism has much to add to this simple-minded conservative approach, given the little information available in the early stages of a new drug's development.

Spiegelhalter and colleagues must be congratulated for making the case for Bayesian approaches to clinical trials so clearly. But, for all their ingenuity, I can only feel that the practical use of Bayesian approaches in medicine will continue to be severely limited for quite some time by the absence of reliable prior information on treatment effects. The very reason to perform a clinical trial is in many cases to show that prior opinions are well intended but misguided—and that as such they may do more harm than good to the interpretation of valuable clinical data.

Bradley P. Carlin (University of Minnesota, Minneapolis): My heartiest congratulations go to the authors on an outstanding paper, which shows the current state of the art in Bayesian design, monitoring and analysis of clinical trials. The authors have been intimately involved in the development of much of the theory and practice described here, and as such this paper represents a crowning achievement.

Despite its philosophical, implementational and documentary advantages, Figs 1(c) and 1(d) illustrate quite clearly why many practitioners (and granting agencies) have remained reluctant to embrace the Bayesian approach: the stopping decision and corresponding trial results may be quite sensitive to the choice of prior distribution. The type of overenthusiasm exhibited by the clinical experts for the potential benefit of neutron therapy is quite common in such settings (see for example Carlin *et al.* (1994)). This dependence is a blessing when good additional prior information is available to facilitate earlier stopping, but a curse when erroneous opinions lead us astray. Clearly the authors' suggestion of simultaneous usage of a broad range of priors (clinical, sceptical and minimally informative) is an important tool in the Bayesian kit, as is the approach in Section 6.3 for assessing prior–data compatibility. Still, the authors 'deliberately do not specify the action to take on observing substantial conflict between prior and data', saying that it must be based on 'individual circumstances'. Can further general guidance be given to assist the decision maker in these equivocal settings?

A partial answer to this question was given by Carlin and Louis (1994), who attempted to characterize the class of priors leading to a given decision (e.g. stopping the trial and deciding in favour of the new treatment) conditional on the observed data. The results are of modest utility under a completely arbitrary prior but can be fairly specific after restricting to a particular parametric form (e.g. the normal family). The approach is in the same pragmatic, data analytic spirit encouraged by the authors of the present paper and can facilitate speculative statements such as 'Given the data accumulated so far, the prior would have to place a mass of at least p on the range where the new treatment is considered superior in order to avoid stopping now and rejecting this treatment as inferior'. Although such statements should never serve as a substitute for results based on careful prior elicitation, they may help a data safety and monitoring board member to reach a decision in settings where the data and his prior are in conflict.

David L. DeMets (University of Wisconsin, Madison) **and K. K. Gordon Lan** (George Washington University, Seattle): The authors are to be congratulated for an excellent and clear presentation of one Bayesian approach to the design, monitoring and analysis of a randomized control clinical trial. On the basis of numerous trials over the past two decades, we believe that no single statistical approach is totally adequate to address the complex decision-making process to terminate trials for convincing evidence of benefit, harm or futility. We believe, as do the authors, that statistical methods can be very useful in assisting that decision process, despite their limitations, if used as guides and not formal rules.

Even though the frequentist approach is well established and widely used, many ideas advocated by Bayesians are already being used implicitly in the design and data monitoring of clinical trials. Unfortunately, the details are not always so well documented. The conditional power approach is one such example. The conditional power under H_0 is evaluated only for the control of α. However, the statistician also provides the conditional power values under various simple alternatives to the data monitoring board (DMB) members at the DMB meetings. Then each DMB member, depending on his or her 'posterior belief', will make a rough weighted average of the conditional power values to form his or her own 'predictive power' for data monitoring and decision-making. This posterior belief of a DMB member is formed by his or her own 'prior' (a physician may have a prior that is different from a statistician's), combined with the data observed in the trial and the external information which became known only after the current trial was initiated. Note that some of the external information known to a DMB member may not be shared by all the DMB members owing to confidentiality. More importantly, most, if not all, DMB members do not use the Bayes formula to pool all the above information forming their posterior belief. A clinical trial may start with a strong belief (based on certain biological theory) that the new treatment would be beneficial. If the accumulated data indicate otherwise, the clinicians start to search for a new biological theory to interpret the observed data. The process described above carries the Bayesian spirit, but it contradicts the formulation of the posterior distribution in a classical Bayesian approach. This approach was used for example in cardiovascular trials such as the beta-blocker heart attack trial and the cardiac arrhythmia suppression trial.

A. P. Dempster (Harvard University, Cambridge): The paper by Spiegelhalter, Freedman and Parmar exhibits pragmatic Bayesianism as welcome commonsense reasoning about uncertainty. Their pragmatism also whittles down details by making assumptions that focus the inference task on a single parameter and finessing other factors by clever devices like ranges of equivalence. The effort in effect seeks a back door into practice by arguing that methods currently based on awkward frequentist logic can be justified through shelf priors characterized as ranging from sceptical to enthusiastic. The examples are accordingly retrospective. Future papers need to remove frequentist crutches, and to rely on Bayes from the start.

My experience with attempting to introduce Bayesian thinking into a major area of practice (Dempster and Hwang, 1993) has been sobering. The various establishments that function as keepers of statistical standards do not recognize overdue needs for serious revisions of practice. When Bayesian statistics eventually achieves its deserved place in a major area of statistical practice such as medical clinical trials, along with other key technologies like transparent data management and access systems, and methodology for applied stochastic modelling, some core features only dimly visible in the current Bayesian literature will surely assume prominence. To their credit, the authors stress the principle that prior knowledge in the form of Bayesian priors be taken seriously throughout the design, monitoring and final analysis of trials and systems of trials, but too much confused wishful thinking, for example about elicitation and communities of priors, surrounds the assembling of evidence necessary to putting priors on the same basis of scientific credibility as technologies that are well established, such as randomization theory and population modelling.

I believe also that there will need to be a radical trend towards formal representation of the real complexities of phenomena, running exactly counter to the striving for simplicity that the paper features. Whereas the Bayesian paradigm can provide realistic assessments of uncertainty in specific empirical situations, wholesale trickery that removes nuisance parameters can only redound in the end to harming the Bayesian enterprise. Reality includes observed survival curves, many relevant covariates and subgroups, parallel end points on side-effects and so forth. If data from trials are not to be wasted on a grand scale, as now, and if simplistic reliance on a few p-values is to be upgraded to more relevant posterior assessments, complex modelling cannot be avoided.

F. D. J. Dunstan (University of Wales College of Medicine, Cardiff): This paper makes a very important contribution to the debate on the analysis of randomized trials and I hope that it will encourage

investigators to make full use of all the information available. The traditional power calculation is open to abuse through a grossly optimistic choice of δ_A—I strongly suspect that the choice is often made simply to give a satisfactory power at a convenient sample size—and an approach which moves away from this is to be greatly welcomed. Like many others, I suspect, my concern is with the choice of prior. Working in an environment where advice is constantly being sought on many projects, it will be difficult always to find time to do a thorough search of information on earlier studies and I suspect that clinical advice will tend to correspond very closely to the enthusiastic prior.

I would be interested to hear the authors' comments in more detail on the recommended prior when previous evidence comes from a non-randomized study. I was recently involved in advising on a protocol submission for a trial involving a new type of post-operative care following a particular form of major surgery. There was a large amount of information on the existing regime and on about 80 cases following a similar operation but with the proposed new pattern of care. These 80 had not been randomized and little information was available on them, although the clinicians were confident that there had been no selection bias. The data suggested that the reduction in length of stay—the end point of the trial—was about 7 days. My approach was to suppose that these findings were possibly biased and unduly optimistic, and to create a prior with this 'treatment effect' of 7 days as a high percentile, say the 95th. It seemed very unlikely to all those concerned that the new care regime would be worse and so 0 was taken to be a low percentile, say the 10th. This gave a prior of mean 3 and standard deviation 2.4. I feel that this is more realistic than taking a mean of about 7 with a large variance, as I think is implied in the paper. Incidentally the approach was rejected by the clinicians because they felt that a Bayesian approach carried too high a risk of being rejected by statisticians on ethical and grant awarding committees. This is a very serious problem.

Susan S. Ellenberg (Food and Drug Administration, Rockville): Spiegelhalter, Freedman and Parmar have performed a great service in contributing to the statistical literature a clear and comprehensive rationale for the increased use of Bayesian methods in designing and assessing the results of clinical trials. They make a very strong case for the value of such methods in improving and enhancing the interpretability of trial results. Since even the most dyed-in-the-wool frequentist will usually admit to at least an intuitive reliance on an informal Bayesian approach to the assessment (as opposed to the analysis) of clinical trials results, there is no doubt that the statistical tools described in this paper could provide much additional insight into clinical trials results if their application were made widespread.

As a statistician working in a regulatory agency, it is of particular interest to me to consider the possible applications of these methods to the regulatory decision-making process. One obvious question is: how will the priors be selected and who will select them? A pharmaceutical sponsor will, by the time that a product has successfully proceeded through all the preliminary phases of testing, have a fairly enthusiastic prior probability for the efficacy of its product. The sincerity of this enthusiasm aside, it is clearly true that, the more enthusiastic the prior, the easier it will be for the sponsor to achieve a high posterior probability of product benefit. The concept of developing a prior by averaging the effects in prior studies, as suggested by the authors, will not usually be applicable to an investigational product.

The authors suggest that regulatory authorities might generally rely on sceptical priors and further suggest that such a prior might be set to be equivalent to having obtained some fraction of the total information in which a zero treatment difference was observed. Such an approach has some intuitive appeal, although it does not seem any less arbitrary than the hypothesis-based procedures for evaluating therapeutic efficacy. It is not difficult to imagine the intensive discussion that might ensue between regulators and sponsors concerning the size of the 'handicap' to be placed on the evaluation of a particular trial! Notwithstanding this concern, the approaches proposed in this paper could be very useful in enhancing our interpretation of trial data; although I am *sceptical* that they will replace current methods, I am *enthusiastic* about adding them to our analytical armamentarium.

Peter Fayers (Medical Research Council Cancer Trials Office, Cambridge): One of the earliest trials in which the authors used the approaches they describe in this paper was the Medical Research Council trial of surgery for gastric carcinoma, comparing standard western surgery (R1) against the more aggressive Japanese surgery (R2). This trial was launched in 1986. 26 surgeons were either interviewed or were sent a questionnaire. They were asked about their 'range of equivalence', their 'prior distribution' for the expected difference in 5-year survival rates and their estimates of base-line (R1) 5-year survival. The results obtained supported the planned sample size of 400 patients. This would enable detection

of a change from 20% 5-year survival (R1) to 33% survival (R2), with 90% power and 5% significance level.

The recruitment of these 400 patients was completed in 1993. While the study was in progress, all information concerning survival rates was kept strictly confidential. In any case, estimates of 5-year survival remain uncertain since few patients were entered more than 5 years ago.

However, over the years it had become apparent that the surgeons' expectations had mellowed. In 1993, before disclosing any survival information, the same questionnaire was used to elicit revised 'prior' beliefs. This confirmed that the surgeons now anticipated a higher base-line (R1) survival rate (27%), and also that their expectations had approximately halved (8% improvement, to 35% for R2). Re-estimating the sample size requirements suggests that 1200 patients would be more appropriate—three times the original study size.

Cancer therapy trials which seek survival improvement often require large sample sizes, are multicentre and take several years to recruit patients. Clinicians frequently start with extreme optimism; that is why they are willing to make so much effort initiating such trials. Later, invariably, a greater sense of realism sets in.

Should we adjust the sample size as 'experts' modify their views? We might expect the revised estimates to be more realistic than the original. Also, such a strategy might make funding easier to obtain: we could design a trial seeking very large effects (fewer patients, less expensive), and later double or treble the sample size!

However, perhaps the clinicians are the wrong people to ask. There is now a long history of cancer clinical trials. Experienced trialists know that very few cancer trials, irrespective of initial euphoria, have found treatment differences of more than a few per cent. For such trials, should we not be using priors strongly centred around 0, irrespective of initial opinions, beliefs and hopes of clinicians?

Mitchell H. Gail (National Cancer Institute, Bethesda): I congratulate the authors on their paper. It illustrates the utility of Bayesian methods in solving practical aspects of clinical trial planning, monitoring and analysis.

There are several reasons why this presentation is appealing. First, the setting out of trial objectives in terms of ranges of clinical inferiority, equivalence and superiority is more realistic than the usual frequentist dichotomization of the parameter space. Second, the authors advocate a conservative approach to monitoring. They avoid premature stopping of trials with early promising results by considering a 'sceptical' prior and premature stopping of trials with early discouraging results by considering an 'enthusiastic' prior. Third, they discuss the frequentist properties of Bayesian procedures. Finally, they advocate an 'explicit separation of the results of a trial from their interpretation'. Thus traditional summary statistics and perhaps likelihood functions would be described in a 'results' section, whereas the interpretation based on various priors would be discussed separately.

However, details are important, and there is room for future work. Obtaining consensus on what constitutes a sceptical prior or a range of treatment 'superiority' will depend on the context and application. Unless some conventions become widely acceptable, these requirements may inhibit widespread use of these methods. Furthermore, this paper emphasizes a well-defined univariate response. Taking into account toxicity or comparing entire survival curves, rather than relying only on the summary hazard ratio, for example, will require an extension of formal methods or other careful, if informal, assessment.

Stephen L. George (Duke University Medical Center, Durham): The authors are to be congratulated for their pragmatic, non-ideological approach and their lucid description of some Bayesian approaches in the design and analysis of clinical trials. One of the reasons for resistance to Bayesian ideas in practice has been the lack of such eloquent advocates who are willing to adopt this approach.

There are two issues raised in this paper on which I would like to comment.

Ethical issues

In discussing example 2 (CHART, Section 4.3) the authors state that 'the ethical basis for the randomization is clear, since the prior probabilities of being on either side of the range of equivalence are almost equal . . .'. A Bayesian approach can shed light on ethical issues, including the ethical judgment that permits randomization of patients, but the ethical basis for individual clinicians is not so clear when, as in this example, both the range of equivalence (δ_I, δ_S) and the prior $p_0(\delta)$ are based on pooling. Only clinician 6 expressed approximately equal prior probabilities of being below or above the

pooled range of equivalence, and this is also the only clinician with an individual prior similar to the pooled prior. Clinician 1 expressed a prior probability of 1.0 that $\delta < \delta_I$ and clinician 9 expressed a prior probability of greater than 0.9 that $\delta > \delta_S$. These two clinicians do not appear to have 'reasonable uncertainty' about the most appropriate treatment and, by implication, may reasonably conclude that it is unethical for them to enter patients.

Predictive distributions

Another application of predictive distributions is the retrospective evaluation of an early stopping decision for trials in which further follow-up of patients who have yet to 'fail' is informative. A decision to stop a trial at some time in the past, after n_1 of n planned patients had been entered and d_1 had failed, can be evaluated by using the current predictive distribution based on follow-up of the $n_1 - d_1$ censored patients to the present. This enables us to use the predictive distribution based on current information to assess what would have happened had we continued the study and entered the remaining $n - n_1$ patients. An example of the use of the predictive distribution in this way is given in George *et al.* (1994). An additional 5 years of follow-up after stopping a trial in non-small-cell lung cancer indicated that the decision to stop after 155 of 240 planned patients were entered was, in retrospect, a reasonable decision, although this reassuring conclusion is not guaranteed in other examples.

Joel B. Greenhouse and Larry A. Wasserman (Carnegie Mellon University, Pittsburgh): We would like to add our congratulations to the authors for their very fine discussion and illustration of the use of Bayesian methods in clinical trials. This paper is distinguished by a high regard for good statistical practice, and as such the authors have successfully advanced the case for the usefulness of Bayesian methods.

Recently we have been investigating an approach to monitoring clinical trials based on the use of robust Bayesian methods. Instead of specifying a single prior distribution or even trying several different prior distributions as suggested by the authors, the robust Bayesian approach to data analysis replaces the prior distribution with a *class* of prior distributions Γ. The aim of this approach is to see how sensitive inferences are to the choice of prior by studying how the inferences might change as the prior varies over this class (Berger, 1990).

A common way to choose Γ is to begin with a specific prior π and then to find a neighbourhood of priors Γ around π. The specific prior as discussed by the authors may be either elicited, 'sceptical', 'enthusiastic' or reference. Since we may not be confident about the choice of prior we consider a class of priors that are similar to π. A tractable class is the *ϵ-contaminated class* defined by $\Gamma_\epsilon = \{(1-\epsilon)\pi + \epsilon Q; Q \in \mathcal{Q}\}$ where $\epsilon \in [0, 1]$ and $\mathcal{Q}$ is the class of reasonable alternative priors. Surprisingly, an easy class to work with is $\mathcal{Q} = \mathcal{P}$, the set of all priors. In practice we compute bounds for the posterior quantity of interest over Γ_ϵ for values of ϵ from 0 to 1, and then plot the bounds *versus* ϵ (see Greenhouse and Wasserman (1994) for details).

We have applied these methods to the authors' analysis of the neutron therapy trial (example 1). We take π to be the 'enthusiastic' prior that they elicited from a group of clinical experts. Fig. 7 is a plot of upper and lower bounds for $P(\delta > \delta_s | x_m)$, $\delta_s = 0.355$, the posterior quantity of interest, as a

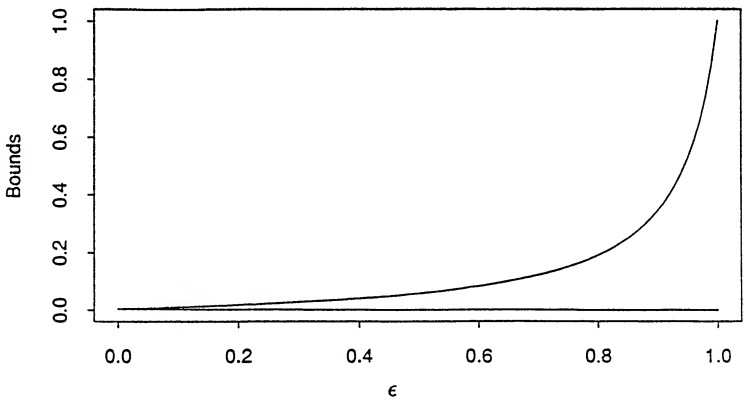

Fig. 7. Bounds on $P(\delta > 0.355 | x_m)$

function of ϵ for the family Γ_ϵ with $\mathscr{Q}=\mathscr{P}$. We find it useful to interpret ϵ as a weight denoting our uncertainty in the specified prior. It is clear from Fig. 7 that even for moderately large amounts of uncertainty in the expert's enthusiastic prior (e.g. $\epsilon=0.50$) the interval between the upper and lower bounds for the posterior quantity of interest is small. Thus, the conclusion that there is small probability that the effect of neutron therapy exceeds the minimal clinically worthwhile benefit of δ_s is quite robust for the prior specification.

Frank E. Harrell, Jr (Duke University Medical Center, Durham): The excellent paper by Spiegelhalter, Freedman and Parmar has both substance and style, quintessential elements of a paper that attempts to change the *status quo* in the midst of historical controversy. As a convicted frequentist, I find that the problem solving tone of this paper, as well as the paper by Hughes (1993), is more convincing than ideological papers in which authors preach 'from the mountain-top'. I expected to disagree with the authors, having to point out dangers in the way that investigators can manipulate Bayesian methods for monitoring trials, but they dealt with the problems forthrightly and proposed useful solutions. I like their use of sceptical priors but remain somewhat sceptical of the use of non-sceptical priors from 'experts' in reporting final results.

The authors have done a real service by supplying free, easy-to-use software for carrying out Bayesian analyses of randomized trials. We have already used their software to great advantage.

Lawrence Joseph and Stanley H. Shapiro (McGill University, Montreal): Although the assumption of a normal distribution with known variance helps to focus attention on the main issues, it is important to note that implementation of the paper's ideas and methods is feasible in the more complex settings encountered in many clinical trials because of the substantial progress in Bayesian computing algorithms.

We strongly endorse the notion of the range of equivalence and wish to emphasize that its utility is independent of whether one adopts a Bayesian approach. By itself, it can go a long way to fostering more clinically relevant designs. It is also useful to consider a two-dimensional range of equivalence. In some instances increased side-effects will accompany increased survival. One compensates for the side-effects by increasing δ_I and/or δ_S appropriately, but the change depends on the (unknown) magnitude of side-effects. A two-dimensional region permits specification of the increase in survival that is necessary for clinical equivalence or superiority for a given level of side-effects. Thus one creates three regions in the plane, rather than on the line. The posterior probability contained within each region can be used to make decisions as before.

Although we are cautioned that the prior may not be sufficiently sceptical if the choice of δ_A is unrealistically large, there is the parallel concern that it may be too sceptical for a prudent choice of δ_A. Formulation of the sceptical prior in the manner indicated in Section 4.1.3 may overpenalize those who plan trials well. Hopefully we shall see increasing numbers of well-designed large trials and a handicap of roughly a quarter ($\beta=0.1$) or a third ($\beta=0.2$) the targeted sample size might be excessive. The issue is not simple and further consideration may lead to refinement of the choice of an appropriate degree of scepticism in different situations.

Most clinical trials have more than one end point of interest. Thus it might be preferable to base sample size calculations on ensuring a sufficiently narrow expected posterior interval for each important parameter, rather than on expected power for a specific critical value.

The phrase 'grouped version of the uncertainty principle' seems somewhat contradictory. The underlying concept, the absence of a consensus in the expert clinical community regarding the comparative merits of the treatments under investigation, seems to us to describe precisely the situation that Freedman (1987) has labelled 'clinical equipoise'.

We shall leave additional comments such as the best means of combining the individual prior distributions to others, and merely wish to close by thanking the authors for this excellent exposition of the Bayesian approach to clinical trials.

Joseph B. Kadane (Carnegie Mellon University, Pittsburgh): This is an excellent paper. What I find so interesting is the collation of more than a decade of pioneering work in applying Bayesian ideas to the analysis of clinical trials. I have a few questions about some of the emphases.

Early in the paper the authors endorse randomization without defining what they mean by it. To some, randomization means the literal application of the binomial or multinomial distribution to the assignment of treatments. To others, it can include assignment for balance among covariates in a way that involves leaving little or nothing to chance. If, for distinction, we call the latter 'controlled' and the

former 'randomized', I hope that the authors will clarify which concept they endorse. (I have elsewhere argued that control is essential but randomization is not (Kadane and Seidenfeld, 1990).)

I am slightly confused about the authors' attitudes towards utilities. On the one hand, they think that the consequences of decisions are 'so uncertain that they make the meaningful specification of utilities rather speculative'. On the other hand, at the end of Section 5.1 they seem to indicate a possibly useful role for utilities. Do they mean their remarks to indicate that they recommend abandoning all hope of explicating values with utilities or that they see utility elicitation as a difficult problem at present not resolved to their satisfaction?

I am unsure about the usefulness of the proposed 'zone of indifference', in which it is not possible to make a decision between the true treatments. For example, if we take literally the statement of Fleming and Watelet cited in Section 7.1.2, there is only the question of whether δ is greater or less than 0.29. If δ is less, the treatment is not sufficiently better than the control to justify the inconvenience, toxicity and expense. Thus would the authors explain more about why they think three regions are needed, in general or in this example?

Although they emphasize the usefulness of Bayesian analysis of data from trials, they might also consider whether there is usefulness for Bayesian ideas at the design stage as well. It is no longer necessary to model the prior of the person using the data as being the same as that of the person(s) analysing the data (Lindley and Singpurwalla, 1991; Etzioni and Kadane, 1993). Also several elicited opinions can be used to protect patients from the most damaging treatments that randomization might force on them, as suggested in Kadane and Sedransk (1980) and implemented in the trial described in Kadane (1986). Would the authors say more about how they expect Bayesian ideas to be useful in the design of clinical trials?

Harold P. Lehmann (Johns Hopkins School of Medicine, Baltimore): Spiegelhalter, Freedman and Parmar are to be congratulated for their synthesis and for making Bayesian methods readily accessible and usable by practising statisticians. I wish to address the authors' interest in the communication between 'statisticians and their clinical colleagues' (Section 8.6).

Dissemination of the results of clinical trials (Section 6.2) is crucially important. Although many non-statistical issues take precedence in a physician's response to such results (Greer, 1988), clinicians are generally concerned with three questions (Sackett *et al.*, 1985).

(a) Are the conclusions clinically meaningful?
(b) Are there threats to internal validity?
(c) Is the population like mine?

The Bayesian framework will be more accessible to readers as more papers are put into an electronic format; novel electronic formats will allow just the sort of interaction that a Bayesian framework requires and that readers require in answering their three basic questions.

Although Spiegelhalter and colleagues reject decision theory as the basis for performing a statistical analysis, decision analysis is the only formal approach that is available to clinicians who wish to evaluate the clinical significance of the results of a clinical trial. Hilden and Habbema (1990) offered a formal framework within which to answer the question; Dr Steve Goodman and I are currently researching the practical application of their framework. The formalism boils down to making explicit the trade-offs in the choices of δ_A (Section 2.2). Explicit modelling of these trade-offs allows the reader to tailor the conclusion to a particular patient or class of patients.

A major strength of Bayesian analysis is its ability to *debias the likelihood function* (an expression I owe to Dr Spiegelhalter), and such debiasing is precisely what is necessary for clinical colleagues to answer questions regarding internal validity. Eddy *et al.* (1992) used such debiasing models in their work in Bayesian meta-analysis. Lehmann and Shachter (1994) used such models to allow physician readers to answer directly this question of study data.

Determining the match of patient populations and the applicability of study results are the most difficult—yet most important—tasks. I suspect that a Bayesian reformulation of the process of randomization would help. Despite the claim by many Bayesians that randomization is irrelevant for inference (Howson and Urbach, 1989), randomization clearly limits cheating and may help with generalization—if we only understood how.

Ettore Marubini and Luigi Mariani (University of Milan): The paper advocates the use of Bayesian methods in the analysis of randomized clinical trials (RCTs) on the grounds of two practical considerations. When interpreting the results of a trial,

(a) the benefits of the therapy investigated must be weighted against its risks, costs and inconveniences, and
(b) prior opinion and evidence internal and external to the study need to be combined.

With reference to the first point, the use of a 'range of equivalence' is suggested. However, such a range could also be dealt with in the context of a classical frequentist analysis.

The second aspect is undoubtedly a more compelling reason for adopting Bayesian methods, since they allow the two sources of evidence (or prior subjective opinion) to be combined in a quantitative rather than informal way.

To keep the discussion on practical ground, one might wonder whether such a formal approach would currently improve the impact of RCT results on the definition of therapeutic strategies. Even though the methodology for performing RCTs is established, many reported studies are questionable because of major flaws in the design or conduct of the study. No statistical methodology can counteract these weaknesses in the analysis phase.

Secondly, RCTs are often criticized because of the need to recruit large samples. The use of Bayesian methods as suggested by the authors fits the framework of falsification that is typical of RCTs and more in general of scientific experiments. Indeed, these methods are suggested as a means to allow for the uncertainty in clinical opinion (and required for ethical reasons) before the study is undertaken. This policy, however, leads to the need for samples that are even larger than those required for a frequentist approach.

Finally, it is stated that one of the aims of the paper is to familiarize the biostatistician with the methods underlying the Bayesian approach. However, statisticians who are willing to shift from the frequentist to a Bayesian framework should be conversant with Bayesian methods not only for the analysis of RCTs but also in other contexts, where statistical tools that are more complex than those described here might be required.

In conclusion, though we basically agree with the authors, we think that these aspects might hamper the use of Bayesian methods or proscribe their practical relevance in the particular context under discussion.

Joseph L. Pater (Queen's University, Kingston): I was pleased to find in this paper a tying together of strands of thought which I have been fascinated by for many years. I agree with the authors' primary argument that context must be considered in planning and interpreting clinical trials, but their analysis of the 5-fluorouracil–levamisole data illustrates how important and difficult the choice of the appropriate perspective is.

As presented, the key issue is to determine to what extent the positive results of the Moertel study should be 'shrunk' by a Bayesian analysis. Following their own advice to use a sceptical prior the authors conclude that the trial results do not provide sufficient evidence to establish 5-fluorouracil–levamisole as clinically superior to the control and that the decision to stop the trial may have been in error. This conclusion is not altered by taking into consideration a previous trial by Laurie which had documented a statistically significant, but less dramatic, benefit from the same regimen.

From the perspective of someone who followed these events closely this formulation misses some of the relevant context. The Moertel trial was a sceptical response to the initial results of the Laurie trial. It was only when an analysis of the accumulating results of the already closed Moertel trial demonstrated a highly significant benefit to 5-fluorouracil–levamisole that the Laurie study was published, followed shortly by the Moertel study. Thus, from a North American perspective the Moertel results should be given additional credit because they arose in the context of attempting to confirm a previous positive study. To me the European scepticism about the conclusions of this trial is due, not to a proper discounting of the findings down into the range of equivalence, but to the manner in which knowledge of these studies became public which in turn influenced the perspective from which the Moertel study was judged. I share the North American view, but the key point is that the context from which a Bayesian analysis is applied can be strongly influenced by how close one is to the actual situation. If, therefore, the context must be considered in judging the results of a new study, it becomes very important to investigate that context as critically as the results of the study itself.

L. I. Pettit (Goldsmiths' College, London) **and K. D. S. Young** (University of Surrey, Guildford): In Section 6.3 the authors discuss the checking of prior and data compatibility. They use Box's (1980) procedure and find a p-value of 'only' 0.14 by using the clinician's prior. Do they consider a p-

TABLE 5

n_0	B_{PN}^{SS}	B_{PN}^{OH}
24	2.03	2.12
48	1.68	1.75
96	1.27	1.33
192	0.98	1.02
$\rightarrow\infty$	0.62	0.64

value of 0.14 'indicative of substantial conflict'?

Young and Pettit (1993) have suggested the use of a Bayes factor to check prior and data compatibility. The starting point is the Bayes factor

$$B_{PN} = p(y|\pi_P)/p(y|\pi_N) \tag{18}$$

where π_P is the assumed proper prior and π_N is a 'non-informative' prior. The idea is to normalize $p(y|\pi_P)$ by comparing it with $p(y|\pi_N)$. Of course $p(y|\pi_N)$ is typically improper and depends on an undefined constant. Young and Pettit compared several methods which have been suggested to overcome this problem. Only those of Spiegelhalter and Smith (1982) which uses a minimal imaginary training set and O'Hagan (1995) which uses a fraction of the whole sample for training are satisfactory in this context. It is also found necessary to use the fractional training sample in both the numerator and the denominator of equation (18). Using the authors' notation we find that the Bayes factors are

$$B_{PN}^{SS} = \left(\frac{1+1/n_0}{1/m+1/n_0}\right)^{1/2} \exp\left\{-\frac{(x_m-\delta_0)^2}{2\sigma^2(1/m+1/n_0)}\right\}$$

using the Spiegelhalter and Smith method and

$$B_{PN}^{OH} = \left(\frac{1+1/n_0}{1/m+1/n_0}\right)^{1/2} \exp\left\{-\frac{(x_m-\delta_0)^2(m-1)/m}{2\sigma^2(1/m+1/n_0)(1+1/n_0)}\right\}$$

using O'Hagan's. These give $B_{PN}^{SS}=5.72$ and $B_{PN}^{OH}=5.73$ for the overview prior and $B_{PN}^{SS}=1.68$ and $B_{PN}^{OH}=1.75$ for the clinician's prior.

Table 5 shows the values of the Bayes factors for the clinician's prior but with varying values of n_0. As n_0 becomes larger there will obviously be more conflict with the data, but with the given values of x_m, δ_0 and σ^2 we must have a very strong prior to find conflict between the data and the prior. Note that as $n_0 \rightarrow \infty$ we are saying that we are certain that the value of δ is 0.169. A classical two-sided test of the hypothesis $\delta = 0.169$ for the data gives a p-value of about 0.025. This would agree with the Bayes factors that there is some evidence for conflict in this extreme case but that it is not overwhelming.

Gary L. Rosner (Duke University Medical Center, Durham): The authors identify many aspects of clinical trial conduct that benefit from a Bayesian approach because it provides a useful paradigm for incorporating external information into the trial's design and analysis. Interim monitoring of clinical trial data is one example. In this paper, the authors propose basing trial termination decisions on the posterior distribution of the parameter characterizing the treatment effect by computing the posterior assuming different prior distributions, each prior specified according to the strength of the implied belief in either the null (i.e. the sceptic's) or the alternative (i.e. the enthusiast's) hypothesis. This has some of the spirit of frequentist uses of the null and alternative hypotheses in significance tests and power calculations, although here one integrates over the parameter space and conditions on the data and either prior distribution. Specifying the precision of the prior and posterior distributions in a way that produces stopping boundaries having good frequentist properties, as discussed here, raises eyebrows among more ideologically directed Bayesian statisticians.

Frequentists, however, walk a tight-rope when directing one to incorporate group sequential boundaries that guarantee the overall type I error into interim monitoring while arguing that crossing these boundaries

should serve as guidelines rather than hard and fast rules. Many frequentists would not adjust the trial's final analyses because of sequential data monitoring, somehow tacitly acknowledging the likelihood principle and ascribing greater weight to it than to inference based on the Neyman–Pearson framework. The data are the data, though; it is the interpretation that may differ from one individual to another. The authors make a useful point by suggesting that one separates out this interpretative part of data analysis from the summary of the trial's quantitative results, pointing out that Bayesian methods incorporate the subjective nature of inference more formally than do frequentist methods, thus allowing individuals to assess the likely treatment benefit given the data in a more straightforward and open way.

The graphical methods for incorporating clinical experts' beliefs and data external to the trial in the prior distribution used when designing and analysing a clinical trial, as well as for communicating clinical trial results and their consistency with different levels of scepticism regarding the benefit of some new therapy, will find appeal among many clinical trialists. I wish to thank the authors for providing a thorough presentation and pragmatic discussion of Bayesian approaches applied to randomized clinical trials.

John Whitehead (University of Reading): This paper is a clear and comprehensive guide to the useful work undertaken by the authors over several years. The suggestions made in Section 8.3 are especially pertinent: quantitative methodology is provided which can supplement the traditional essay which forms the discussion section of a clinical trial report. It is incumbent on investigators to put their findings into the context of prior opinions and the results of others: here are the tools with which this can be done. The authors suggest that the results section presents the data found in the trial being reported. This is also reasonable, but there is no reason why conventional safeguards against overinterpretation—significance levels and confidence intervals—should not be included. My main reservation about the paper is its potential influence on the design of studies and the possibility of leading people away from sample sizes chosen to achieve small error rates.

I am particularly concerned about the Bayesian monitoring criteria described in Section 5.1 of the paper. Although the authors are reluctant to define stopping rules, suppose that a reader of their paper decided to conduct a series of interim analyses and to stop when the subjective distribution for δ satisfied either $P(\delta > 0) < \epsilon$ or $P(\delta < 0) < \epsilon$. Suppose further that the value $\epsilon = 0.025$ was chosen, and that a uniform, non-informative prior was taken for δ. The resulting procedure would be equivalent to conducting a series of conventional significance tests, each at the nominal level $\alpha = 0.05$. This procedure inflates overall error rates, as pointed out by Armitage *et al.* (1969), and yet, by a different route, we are led to contemplate it again. Should an informative prior be used, presumably expressing some belief in the benefit of the experimental treatment, then error rates may be increased further.

It may be because of the small sample sizes and large error rates of Bayesian sequential procedures based on true prior beliefs that the authors have introduced their artificial sceptical and enthusiastic priors expressing, not the investigators' beliefs, but the hypothetical beliefs of hypothetical individuals. These are really tools to slow down the stopping rule, increase sample size and reduce error rates. They are tools which are not chosen according to familiar considerations of frequentist error probabilities or optimization of utility. This is really new territory, unfamiliar from any other field of statistical application. Novelty is not bad in itself; however, clinical investigators and medical statisticians should be cautious before leaving traditional safeguards against error of wrong decisions behind.

H. Sam Wieand (Mayo Clinic, Rochester): This work shares a property of previous papers by the authors in that it very clearly describes some practical applications of Bayesian theory to clinical trials. At two recent international workshops involving cancer co-operative group statisticians, terms such as 'range of equivalence' and 'sceptical priors' were part of the standard vocabulary. This is a clear change from 5 years ago and, to a large degree, reflects the effect that these same authors have had with some of their earlier publications. Acceptance of these Bayesian concepts by co-operative group statisticians is a considerable accomplishment, given the strong commitment that clinical trial statisticians have given to randomization and subsequent frequentist properties.

In the brief space available for a comment, it is difficult to address any technical details, so I shall focus on the discussion in Section 8 which outlines practical considerations. Consideration of frequentist properties of Bayesian monitoring procedures, as in the authors' Table 4, is an important step in bringing Bayesian methods forwards. I would be very uncomfortable using a design if I did not have a sense of the probability of a type I error. There are two primary reasons. The first is the unfortunate fact, noted by the authors, that many chemotherapy treatments do not offer an improvement over no treatment or standard regimens. The second is that many clinical trial statisticians will be responsible for scores

if not hundreds of trials over a career, so the frequentist idea of keeping the number of times that a null hypothesis is falsely rejected is meaningful to us.

However, when listening to or reading presentations of trial results, I have found that discussions using Bayesian concepts, such as sceptical and enthusiastic priors, are very informative and helpful. Thus I would support the authors' suggestion that results be presented in a traditional section, followed later with a discussion using (but not limited to) Bayesian methods. Space limitations in journals may be an issue, but I think that a well-written interpretation section would have a good chance of being accepted, particularly if a phase III trial was being presented.

In summary, Bayesian ideas have already crept into cancer clinical trial discussions and such discussions have led to more clarity in interpreting trial designs and results. I also believe that frequentist concepts such as randomization and type I errors are still fundamental to clinical trials, so an understanding of their practical relationship is quite valuable and I thank the authors for a presentation which helps with that process.

The **authors** replied later, in writing, as follows.

We are grateful and somewhat overwhelmed by the many contributions to the discussion, and in the available space we can only make a limited response to the issues raised. Our reply is divided into four main sections: the first two deal with the controversial issues of the source of priors and the role of type I error, then we consider the problems surrounding the introduction of Bayesian methods into the clinical trial culture and finally we briefly cover some technical points.

Source of priors

Many discussants (Lewis, Buyse, Fayers, Altman, Evans, Senn and Harrell) question the relevance and role of clinical (subjective) prior information as illustrated in the neutron therapy and CHART examples. Some suggest that the neutron therapy example reflects their own (subjective) belief that clinical opinion is often not very useful and that more, if not all, emphasis should be placed on the prior opinion obtained from a meta-analysis of previous studies. Although this suggestion has some obvious appeal, achieving apparently a greater degree of objectivity, such an approach loses sight of many of the complexities in designing and monitoring trials. In particular, at the design stage of this trial, although the participating clinicians were aware of the previous studies, using a variety of different dosage regimens, suggesting that neutron therapy may actually be worse than conventional therapy, they considered that the high energy of the neutrons to be used in this trial were sufficiently different from that used in these previous trials for them to expect that there was a good chance that survival would be materially improved.

In the event, the elicitation of the expected benefit and the range of equivalence allowed the conclusion to be reached that even individuals optimistic towards this high energy level of neutrons were likely to be convinced by the trial data that this treatment was not clinically worthwhile. This result was achieved at a relatively early stage—after 151 patients had been randomized. Of course, the trial results taken together with a meta-analysis of previous trials provide overwhelming evidence against neutron therapy, and given the results of this trial it would seem that there is little difference between the various energies. However, whether these trials should be combined is principally a biological and clinical question and not a statistical one (despite the fact that no obvious statistical heterogeneity was observed). Just because the trial results support the meta-analysis result does not mean that we can combine them.

This example illustrates that both a meta-analysis of previous randomized trials and clinical prior opinion have a place in the design, monitoring and interpretation of trials. Clearly, in situations in which the trial is very similar to previous trials, which themselves are judged to be similar, emphasis should be placed on the meta-analyses prior and posterior. However, when there are no previous randomized trials—such as the CHART trials—or there is perhaps considerable biological and clinical heterogeneity, such as the neutron therapy trials, subjective judgment, sometimes in the form of a clinical prior distribution, should play a major role in the design, monitoring and final interpretation of the trial.

When obtaining subjective clinical opinion in the form of prior distributions and ranges of equivalence, Dr Altman asks whom and how many should we ask. In all our examples we have elicited information from those clinicians intending to randomize patients into the trial. The ethical basis for randomization both at the beginning and during the trial are most relevant to these clinicians. These clinicians are not a random sample of all clinicians treating this disease, but this is not the aim. In our experience these clinicians usually represent a well-informed sample of individuals who are often more enthusiastic than their colleagues towards the new treatment.

Many discussants (Grieve, Pocock, Herson and Altman) comment on our approach of using the

arithmetic average of the nine clinical prior distributions in the CHART example. We acknowledge that there is considerable variability across these nine clinicians. However, in eliciting opinions our aim was to obtain an estimate of a prior distribution of a 'typical well-informed clinician'. Arithmetic pooling, in which each participating clinician's opinion represents a data point, provides this. Such pooling also has the added benefit that it provides a level of variability which is likely to be more representative than many individual prior distributions. This approach does not disqualify the use of individual prior distributions at the termination of the trial to, perhaps, represent more extreme views. Dr Grieve further suggests that rather than to use the sceptical and enthusiastic prior distributions constructed around the alternative hypothesis and zero difference we should use the most sceptical and enthusiastic elicited priors (clinicians 1 and 9 for the CHART example). Dr Ashby also questions the 'classical' frequentist motivation for the sceptical and enthusiastic prior distributions, and enquires about their Bayesian motivation. A frequentist motivation for the sceptical prior is given in Sections 4.1.3 and 8.1. A Bayesian motivation for this prior is given in Section 8.2 and referred to by Mr Fayers, i.e. that in many areas most new treatments are found to be ineffective or at best marginally effective. (This also addresses the point made by Dr Buyse and Dr Machin who both suggest that in most cases we shall have very little information about the size of the treatment effect.) The choice of the enthusiastic prior provides a counterbalance to the sceptical prior, and as we suggest in Section 4.1.4 (and report in Parmar *et al.* (1994)) this formal prior is usually very close to the elicited clinical prior. The sceptical and enthusiastic priors of clinicians 1 and 9 in the CHART example do not have these appealing properties. Professor George suggests that, in fact, for these clinicians, their prior opinions and ranges of equivalence are such that they should perhaps not be entering patients into the CHART trials. We agree.

Several discussants suggest that it is impractical, cumbersome or confusing to use a community of priors. We disagree. With software now available the calculations for a variety of prior distributions is simple to perform. We refer them to the levamisole + 5-fluorouracil and the misonidazole trials examples and to Parmar *et al.* (1994) who show that it is possible to consider and present a range of prior-to-posterior analyses, which aid interpretation of the results. In fact, if the principal aim of the trial is to provide evidence which is convincing to a wide body of opinion, we would argue that such 'sensitivity' analyses are essential to assess likely variations in interpretation.

Professor Bailey and Dr Buyse suggest that there are some perfectly good frequentist solutions to some of the problems that we address. In particular they respectively suggest that we could use two different null hypotheses, instead of sceptical and enthusiastic priors, and a 'simple-minded' conservative approach of two p-values less than 0.05 to reflect scepticism. These methods reflect the problems with many traditional frequentist approaches, emphasizing hypothesis testing and p-values over estimation (both point and interval), and confuse the range of equivalence with the prior distribution. The Bayesian approach provides a framework for providing coherent point and interval estimates with a variety of prior beliefs, together with a facility to make statements about the probability of being above and below the range of equivalence.

Mr Fayers asks whether we should adjust sample sizes as clinicians modify their views through the course of a trial. This is clearly a desirable property since, for example, information from other trials may come to light which influences the conduct of the current trial. The Bayesian approach allows this to be done through incorporation of the new information in the prior distribution. Such interim adjustment is not possible in the traditional frequentist framework.

Monitoring trials and type I error

We find it heartening that an increasing number of statisticians actively involved in clinical trials (e.g. Grieve, Herson and Rosner to name a few) agree with us that the Bayesian approach carries important advantages over the frequentist approach in clinical trials monitoring.

Professor Pocock, Professor Turnbull and Professor Whitehead express concern that a Bayesian approach to monitoring trials will dispense with control of the type I error. We refer them again to our Section 8.1. As Cornfield (1966) pointed out, abiding concern with the type I error must indicate a sceptical attitude towards new treatments. Therefore it is gratifying, but not surprising, that adopting a sceptical prior leads naturally to the control of type I error. See also the forthcoming paper by Grossman *et al.* (1994).

Dr Grieve asks our view on Pitman's result, indicating that it might be more appropriate to use increasing, rather than decreasing, p-values for stopping rules. We have not considered this in detail, but it occurs to us that this result could be a direct consequence of formulating the monitoring problem as a test of two simple hypotheses. This formulation seems unnecessarily artificial, and we prefer to

deal with monitoring as an estimation problem, preserving the full range of possible states of nature.

Professor Armitage suggests that in Fig. 4 situations B and D should not lead to early stopping. We agree that in many trials it would be appropriate to continue in these situations. However, where the data indicated that a new treatment was possibly harmful in the long term, and very unlikely to achieve a long-term benefit sufficient to overcome completely its extra toxicity (corresponding to situation B), we may wish to terminate.

Professor Bather and Professor Turnbull ask whether we recommend the use of defined stopping rules. The answer is yes. Section 8.1 suggests the type of prior distribution(s) that could be used. The range of equivalence adopted would depend on the perceived relative toxicities, inconveniences and costs of the treatments that are being compared. Like others, we intend the stopping rules to act as a starting point for wider discussion regarding the future of the trial. Unlike frequentist rules, the Bayesian approach would not be disturbed if the rules were not followed.

Abrams, Jennison, DeMets and Lan mention other statistical methods of monitoring trials. Dr Abrams mentions stochastic curtailment, and Professor Jennison mentions repeated confidence intervals. These have been debated and contrasted elsewhere (Jennison and Turnbull, 1989, 1990; Freedman *et al.*, 1994). Professor DeMets and Professor Lan describe a procedure that they have used but which is not well documented. It is a modified version of stochastic curtailment, presenting conditional powers for a range of possible treatment effects. The frequentist properties of this procedure are unknown.

Finally Professor Whitehead describes our suggestion to use sceptical priors and posterior probabilities as 'new territory', but this is to deny Cornfield (1966) the rights to introducing the sceptical prior, and Laplace (1774) the rights to introducing the practical application of posterior probability distributions!

Bayesian methods in trial culture

Several discussants raised questions about the role of Bayesian methods in drug regulation. Grieve, Ashby and Ellenberg perceive an increasing willingness on the part of regulatory bodies to deal with Bayesian approaches. We agree that this softening of attitude is encouraging. However, to date, few, if any, pharmaceutical companies have tested the water by employing Bayesian methods of design or analysis. In our experience, the pharmaceutical industry tends to follow the lead of the regulatory bodies and is less likely to change its practices unless the regulatory bodies actively encourage it.

Professor Keiding notes the inevitability that the regulatory bodies will issue defined guidelines for reporting of trials and asks whether these are likely to be any less ridiculous if a Bayesian paradigm were used. We think that there would be some improvements: for example, the regulatory body would no longer have a major concern over the number of interim analyses that are performed during a trial.

Dr Herson, Professor Armitage and Dr Ellenberg comment on the nature of the prior distribution to be used by the regulatory body. We would envisage the use of a sceptical prior. We view the role of the regulatory body as that of public watch-dog, and a degree of scepticism is consistent with such a role. The pharmaceutical industry is already familiar with the regulatory bodies adoption of a sceptical stance, so we do not expect great difficulties here. As Dr Ellenberg notes, there will inevitably arise a certain arbitrariness in specifying the degree of scepticism adopted by the regulatory body, but this will be of the same type as is involved with our current bench-marks of 5% significance and 80% or 90% power.

Professor Armitage questions whether a sceptical prior is appropriate for a regulatory body in the face of an apparently harmful effect of a new treatment. Assuming that the data are a surprise, it seems to us reasonable to adopt a somewhat sceptical attitude towards them. However, with a range of equivalence of $(0, D)$, where D is positive, and allowing termination under situation B of Fig. 4, we would tend to stop more readily in the face of harmful effects than beneficial effects.

Pocock, Healy, Evans and Marubini all question how a re-education of statisticians and clinicians can take place to accommodate Bayesian thinking. We support Evans in favouring a gradual rather than revolutionary approach but agree that Bayesian analyses will tend to be more complex and, heaven forbid, take more time. However, we cannot see the need to *think* about the implication of the results as being a bad development. The discussants illustrate the difficulties of satisfying all opinions: some argue for simple methods, whereas others (e.g. Professor Dempster) claim with some justification that we have oversimplified the models. The huge recent developments in flexible software for more complex Bayesian analysis now mean that we can deal with realistically complex models incorporating, for example, multiple end points (see the remarks by Professor Turnbull and Professor Gail). However, the use of more complex methods can only be introduced in so far as they can be communicated effectively and understood intuitively by medical researchers. We disagree with Professor Pocock that software will

hamper us: the Cox proportional hazards model at first appeared prohibitively computationally complex.

Technical aspects

Jennison, Lindley, Turnbull, Berry and Kadane all argue for a full decision theoretic formulation, and a challenge is made to find whether our stopping procedure has any reasonable coherent justification. We again appeal to our 'gradualist' sentiments and are certainly interested in a coherent approach with explicit utilities. However, it has been enough of a struggle in working to put subjective probability judgments at least on the road to respectability, and utilities appear to us an order of magnitude more difficult. We hope that Professor Lindley will be able to welcome the relevant paper in 25 years' time.

Bather, Bailey, Kadane and Lehmann ask about our justification for requiring randomization. Our reasoning is that randomization is required to ensure that the selection of treatment for each patient is unbiased, i.e. undisturbed by the clinician's propensity to choose one or other treatment for an individual. Such conditions are necessary for the correct specification of the likelihood, and, since the correct calculation of the posterior distribution depends on the likelihood being correct, it follows that Bayesians as well as frequentists require 'randomization' to evaluate treatments in clinical trials properly. By randomization, we mean all allocation methods that avoid biased treatment selection. We would include in this definition deterministic methods such as minimization (Pocock and Simon, 1975).

Unlike Professor Lehmann we do not think that randomization helps with generalization to a wider population of patients, other than its helping towards an unbiased comparison of the two treatments within the sample of patients entering the trial.

Professor Bailey asks whether our Bayesian analysis would be different under different randomization schemes. The answer is perhaps. For example, if randomized blocks were used because of a real expectation that the treatment responses would improve over the course of the trial, then we would want to include the block in our likelihood model. Similarly, if there were covariates that were known to be strongly related to the response, then we would wish to include such covariates in the likelihood model. We might then use a fully Bayesian procedure specifying priors for the block or covariate effects, or we might take a short cut and obtain an 'adjusted' estimate and standard error of treatment effect from a classical analysis and use this as the likelihood in our Bayesian analysis.

Dr Gore asks about adaptive randomization similar to methods proposed by Bather (1985) and others, where the randomization ratio varies according to the latest results from the trial. We have always had some difficulty with this concept. Randomization is justified by continuing uncertainty over which treatment is superior. Changing the randomization ratio is a signal to the public that the evidence is swinging in a particular direction, and seems to be an invitation to interference from outside in the conduct of the trial.

Pocock, Evans, Atkinson and Kadane ask about the use of Bayesian methods in the design of clinical trials. We did not give this area much emphasis in our paper simply because we have spent less time thinking about it! We think that there is a role for Bayesian methods in calculating sample size for clinical trials, using the concept of average power as the criterion to be guaranteed (Spiegelhalter and Freedman, 1986). Our problem in applications is that we are unsure what level of average power should really be required. We also think that there is a definite role for Bayesian methods in planning factorial design trials. There is always debate about the likelihood of interaction when discussing such designs, and placing a prior distribution on the interaction should provide a useful way of dealing with this issue. Professor Atkinson mentions work on Bayesian designs that tend to equalize covariates over treatment groups. Non-Bayesian designs proposed by Pocock and Simon (1975) and by Begg and Iglewicz (1980) have a similar aim; these are often used in practice. A comparison of these methods with the Bayesian designs would be of interest.

Finally, Professor Marubini and Dr Mariani complain that Bayesian methods do not solve the problem of major flaws in the design or conduct of trials, nor do they reduce the sample size requirements. We agree with both these comments. It is very important for statisticians to involve themselves in the substantive medical issues of clinical trials, as this will help them to advise clinicians effectively on good trial design and conduct. On sample size, we fully support the recommendations of Peto and others who advise that trials are still often too small to answer effectively important public health issues that confront us. We think that the proper use of Bayesian methods will endorse this view, together with the emerging realization that $p<0.05$ is an inappropriate and weak criterion for accepting new treatments.

Dr Senn, Professor Joseph, Professor Shapiro and others point out that there is nothing Bayesian about introducing a range of equivalence. We agree, although we emphasize that its use fits in very naturally with a Bayesian analysis that leads to posterior probabilities of the treatment falling below,

within or above this range. We cannot think of a frequentist approach that would lead to such an appealing summary of the inferences, although hypothesis testing about each end of the range can be adopted (Freedman *et al.*, 1984; Fleming and Watelet, 1989).

Professor Kadane questions whether it is useful to specify a range of equivalence. We are sure that Fleming and Watelet (1989) were using this concept, since their stopping rules were to advocate termination either if the treatment difference was positive and significantly different from 0 or if the difference was significantly less than 0.29. Generally, we think that the range of equivalence can be a useful concept when the treatments to be compared carry very different levels of toxicities, difficulties or costs, and where the balancing of such inequities with potential benefits in treatment outcome is not obvious. In these cases it is useful to specify limits beyond which a large proportion of clinicians would agree that the benefits respectively outweigh or are insufficient to outweigh the inequities. We also think that there will be some trials where a point of equivalence (not necessarily zero difference) will suffice.

Professor Joseph and Professor Shapiro raise the possibility of a two-dimensional region of equivalence. This is attractive when there is still considerable uncertainty regarding the levels of toxicity that will occur, so that the trial will provide substantial extra information on toxicity as well as on the main outcome. Professor Peter Thall (M. D. Anderson Hospital, Houston, Texas) is working on these problems from a Bayesian perspective. It is perhaps this two-dimensional approach to which Dr Gore refers when she talks about a 'prior distribution on the range of equivalence'.

Dr Buyse questions the appropriate behaviour when a conflict is found between the data and prior, and we reiterate our opinion that this is a matter for the specific monitoring committee. The explicit recording of a prior is of great value even, and perhaps especially, if it is found to be misguided. Carlin, Young and Pettit provide innovative tools that could be used to help in the discussion of the monitoring committee in such a situation.

Dr Pater points out that the levamisole + 5-fluorouracil example illustrates the need to be aware of the chain of events, as well as the study results themselves. In particular, he points out that the trial reported by Laurie and co-workers was performed before the levamisole + 5-fluorouracil trial that we report, and in fact that this latter trial was performed as a so-called 'confirmatory' trial. We agree. However, the combined analysis including both trials is still valid in the sense that it is combining results from different trials together with sceptical and enthusiastic prior distributions, but perhaps not in the chronological order. Despite this, we believe that this combined analysis does help to explain why the more enthusiastic North American clinicians found the results more convincing than their sceptical European colleagues.

Professor Armitage refers to the 1993 National Institutes of Health Revitalization Bill (that has since become an Act) which has a section referring to the inclusion of women and minorities as participants in clinical research projects. We refer Professor Armitage to the Institutes' guidelines on the implementation of this Act (Federal Register, 1994) which, in fact, follow quite closely in the spirit of Professor Armitage's suggestion of adopting a multivariate prior assuming high correlation between treatment effects for different strata, although the discussion is not couched in such formal terms!

Finally we again take the opportunity of thanking all the discussants for their contributions.

REFERENCES IN THE DISCUSSION

Abrams, K. R., Ashby, D. and Errington, R. D. (1994) Bayesian parametric survival models—an application to a cancer clinical trial. Submitted to *Statist. Med.*

Anscombe, F. J. (1963) Sequential medical trials. *J. Am. Statist. Ass.*, **58**, 365–383.

Armitage, P., McPherson, C. K. and Rowe, B. C. (1969) Repeated significance tests on accumulating data. *J. R. Statist. Soc.* A, **132**, 235–244.

Atkinson, A. C. (1982) Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, **69**, 61–67.

Atkinson, A. C. and Donev, A. N. (1992) *Optimum Experimental Designs*. Oxford: Clarendon.

Bather, J. A. (1985) On the allocation of treatments in sequential medical trials. *Int. Statist. Rev.*, **53**, 1–13.

Begg, C. B. and Iglewicz, B. (1980) A treatment allocation procedure for sequential clinical trials. *Biometrics*, **36**, 81–90.

Berger, J. (1990) Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Planng Inf.*, **25**, 303–328.

Berry, D. A., Wolff, M. C. and Sack, D. (1992) Public health decision making: a sequential vaccine trial (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 79–96. Oxford: Oxford University Press.

——(1994) Decision making during a phase III randomized controlled trial. *Contr. Clin. Trials*, **15**, in the press.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A, **143**, 383–430.

Box, G. E. P. and Lucas, H. L. (1959) Design of experiments in nonlinear situations. *Biometrika*, **46**, 77–90.
Brown, L. D., Cohen, A. and Strawderman, W. E. (1980) Complete classes for sequential tests of hypotheses. *Ann. Statist.*, **8**, 377–398.
Browner, W. S. and Newman, T. B. (1986) Confidence intervals. *Ann. Intern. Med.*, **105**, 973–974.
Carlin, B. P., Chaloner, K. M., Louis, T. A. and Rhame, F. S. (1994) Elicitation, monitoring, and analysis for an AIDS clinical trial (with discussion). In *Case Studies in Bayesian Statistics*, vol. 2. New York: Springer. In the press.
Carlin, B. P. and Louis, T. A. (1994) Identifying prior distributions that produce specific decisions, with application to monitoring clinical trials. In *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner* (eds D. Berry, K. Chaloner and J. Geweke). Amsterdam: North-Holland. In the press.
Chaloner, K. and Larntz, K. (1989) Optimal Bayesian design applied to logistic regression experiments. *J. Statist. Planng Inf.*, **21**, 191–208.
Colton, T. (1963) A model for selecting one of two medical treatments. *J. Am. Statist. Ass.*, **58**, 388–400.
Cornfield, J. (1966) Sequential trials, sequential analysis and the likelihood principle. *Am. Statistn*, **20**, 18–23.
Davenas, E., Beauvais, F., Amara, J., Oberaum, M., Robinson, B., Miadonna, A., Tedesche, A., Tomeranz, B., Fortner, T., Belon, P., Saint-Laudy, J., Poitevin, B. and Benvenista, J. (1988) Human basophil degranulation is triggered by very dilute antiserum against human IgE. *Nature*, **333**, 816–818.
DeMets, D. L. (1984) Stopping guidelines vs stopping rules: a practitioner's point of view. *Communs Statist. Theory Meth.*, **7**, 389–398.
Dempster, A. P. and Hwang, J.-S. (1993) Component models and Bayesian technology for estimation of state employment and unemployment rates. In *Proc. 1993 A. Res. Conf.*, pp. 571–581. Washington DC: US Bureau of the Census.
Eales, J. D. and Jennison, C. (1992) An improved method for deriving optimal one-sided group sequential tests. *Biometrika*, **79**, 13–24.
Early Breast Cancer Trialists' Collaborative Group (1992) Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. *Lancet*, **339**, 1–15, 71–85.
Eddy, D. M., Hasselblad, V. and Shachter, R. (1992) *The Statistical Synthesis of Evidence: Meta-analysis by the Confidence Profile Method.* Boston: Academic Press.
Efron, B. (1971) Forcing a sequential experiment to be balanced. *Biometrika*, **58**, 403–417.
Etzioni, R. and Kadane, J. B. (1993) Optimal experimental design for another's analysis. *J. Am. Statist. Ass.*, **88**, 1404–1411.
European Committee for Proprietary Medicinal Products (1993) *Note for Guidance: Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medicinal Products (Draft 4, March 1993).* Brussels: European Committee for Proprietary Medicinal Products.
Fayers, P. M. and Armitage, T. (1993) Towards an international register of cancer trials: the UKCCCR Register of UK Trials. *Eur. J. Cancer* A, **29**, 907–912.
Fayers, P. M., Cook, P. A., Machin, D., Donaldson, N., Whitehead, J., Ritchie, A., Oliver, R. T. D. and Yuen, P. (1994) On the development of the Medical Research Council trial of α-interferon in metastatic renal carcinoma. *Statist. Med.*, **13**, in the press.
Federal Register (1994) NIH guidelines on the inclusion of women and minorities as subjects in clinical research. *Fed. Reg.*, **59**, 14508–14513.
Fisher, R. A. (1950) *Statistical Methods for Research Workers*, 11th edn, p. 20. London: Oliver and Boyd.
——(1960) *The Design of Experiments*, 7th edn. Edinburgh: Oliver and Boyd.
Fleming, T. R. and Watelet, L. F. (1989) Approaches to monitoring clinical trials. *J. Natn. Cancer Inst.*, **81**, 188–193.
Freedman, B. (1987) Equipoise and the ethics of clinical research. *New Engl. J. Med.*, **317**, 141–145.
Freedman, L. S., Lowe, D. and Macaskill, P. (1984) Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, **40**, 575–586.
Freedman, L. S. and Spiegelhalter, D. J. (1983) The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Statistician*, **32**, 153–160.
Freedman, L. S., Spiegelhalter, D. J. and Parmar, M. K. B. (1994) The what, why and how of Bayesian clinical trials monitoring. *Statist. Med.*, to be published.
Freeman, P. R. (1993) The role of *p*-values in analysing trial results. *Statist. Med.*, **12**, 1443–1452.
George, S. L., Li, C., Berry, D. A. and Green, M. R. (1994) Stopping a clinical trial early: frequentist and Bayesian approaches applied to a CALGB trial in non-small-cell lung cancer. *Statist Med.*, to be published.
Gore, S. M. (1994) The consumer principle of randomisation. *Lancet*, **343**, 58.
Greenhouse, J. and Wasserman, L. (1994) Robust Bayesian methods for monitoring clinical trials. *Technical Report 588.* Department of Statistics, Carnegie Mellon University, Pittsburgh.
Greer, A. L. (1988) The state of the art *versus* the state of the science: the diffusion of new medical technologies into practice. *Int. J. Technol. Assmnt Hlth Care*, **4**, 5–26.
Grieve, A. P. (1994) Extending a Bayesian analysis of the two-period crossover to allow for baseline measurements. *Statist. Med.*, **13**, 905–929.
Grossman, J., Parmar, M. K. B., Spiegelhalter, D. J. and Freedman, L. S. (1994) Unified hypothesis testing, point estimation and interval estimation for group sequential clinical trials. *Statist. Med.*, to be published.
Grundy, P. M., Healy, M. J. R. and Rees, D. H. (1956) Economic choice of the amount of experimentation (with discussion). *J. R. Statist. Soc.* B, **18**, 32–55.

Hilden, J. and Habbema, J. D. F. (1990) The marriage of clinical trials and clinical decision science. *Statist. Med.*, **9**, 1243–1257.
Hildreth, C. (1963) Bayesian statisticians and remote clients. *Econometrica*, **31**, 422–438.
Hirst, S. J., Hayes, N. A., Burridge, J., Pearce, F. L. and Foreman, J. C. (1994) Human basophil degranulation is not triggered by very dilute antiserum against human IgE. *Nature*, **366**, 525–527.
Howson, C. and Urbach, P. (1989) *Scientific Reasoning: the Bayesian Approach*. La Salle: Open Court.
Hughes, M. D. (1993) Reporting Bayesian analyses of clinical trials. *Statist. Med.*, **12**, 1651–1663.
International Conference on Harmonisation (1993) *Draft Tripartite Guidelines (Draft 2.0, March 1993): Dose Response Information to Support Drug Registration*.
Jennison, C. and Turnbull, B. W. (1989) Interim analyses: the repeated confidence interval approach (with discussion). *J. R. Statist. Soc.* B, **51**, 305–361.
——— (1990) Statistical approaches to interim monitoring of medical trials: a review and commentary. *Statist. Sci.*, **5**, 299–317.
——— (1993) Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics*, **49**, 741–752.
Kadane, J. B. (1986) Progress toward a more ethical method for clinical trials. *J. Med. Phil.*, **11**, 385–404.
Kadane, J. B. and Sedransk, N. (1980) Toward a more ethical clinical trial. In *Bayesian Statistics* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 329–338. Valencia: Valencia University Press.
Kadane, J. B. and Seidenfeld, T. (1990) Randomization in a Bayesian perspective. *J. Statist. Planng Inf.*, **25**, 329–345.
Lan, K. K. G. and DeMets, D. L. (1983) Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659–663.
Laplace, P. (1774) Memoire sur la probabilité des causes par les évenements. *Mem. Acad. R. Sci. Pres. Div. Sav.*, **6**, 933–937.
Lehmann, H. P. and Shachter, R. D. (1994) A physician-based architecture for the construction and use of statistical models. *Meth. Inform. Med.*, to be published.
Lindley, D. and Singpurwalla, N. (1991) On the evidence needed to reach agreed action between adversaries with application to acceptance sampling. *J. Am. Statist. Ass.*, **86**, 933–937.
Machin, D. (1992) Interim analysis and ethical issues in the conduct of trials. In *Introducing New Treatments for Cancer: Practical, Ethical and Legal Problems* (ed. C. J. Williams). Chichester: Wiley.
Makuch, R. and Johnson, M. (1989) Issues in planning and interpreting active control equivalence studies. *J. Clin. Epidem.*, **42**, 503–511.
Mehring, G. H. (1993) On optimal tests for general interval-hypotheses. *Communs Statist. Theory Meth.*, **22**, 1257–1297.
Numerical Algorithms Group (1983) *NAG Library Introductory Guide*. Oxford: Numerical Algorithms Group.
O'Hagan, A. (1995) Fractional Bayes factors for model comparisons (with discussion). *J. R. Statist. Soc.* B, **57**, 99–138.
Parmar, M. K. B., Spiegelhalter, D. J., Freedman, L. S. and CHART Steering Committee (1994) The CHART trials: Bayesian design and monitoring in practice. *Statist. Med.*, **13**, in the press.
Pocock, S. J. and Simon, R. (1975) Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, **31**, 103–115.
Pocock, S. J. and Spiegelhalter, D. J. (1992) Grampian region early anistreplase trial (GREAT). *Br. Med. J.*, **305**, 1015.
Sackett, D. L., Haynes, R. B. *et al.* (1985) *Clinical Epidemiology—a Basic Science for Clinical Medicine*. Boston: Little, Brown.
Senn, S. J. (1993) Inherent difficulties with active control equivalence studies. *Statist. Med.*, **12**, 2367–2375.
Smithells, R. W. and Sheppard, S. (1980) Letter to the Editor. *Lancet*, i, 647.
Smithells, R. W., Sheppard, S., Schorah, C. J., Seller, M. J., Nevin, N. C., Harris, R., Read, A. P. and Fielding, D. W. (1980) Possible prevention of neural-tube defects by periconceptional vitamin supplementation. *Lancet*, i, 339–340.
Spiegelhalter, D. J. and Freedman, L. S. (1986) A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statist. Med.*, **5**, 1–13.
——— (1988) Bayesian approaches to clinical trials. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 453–477. Oxford: Oxford University Press.
Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. (1993) Applying Bayesian thinking in drug development and clinical trials. *Statist. Med.*, **12**, 1501–1511.
Spiegelhalter, D. J. and Smith, A. F. M. (1982) Bayes factor for linear and log-linear models with vague prior information. *J. R. Statist. Soc.* B, **44**, 377–387.
Statistical Sciences Inc. (1991) *S-PLUS Version 3.0—Reference Manual*, vols 1, 2. Oxford: Statistical Sciences.
Stewart, L. A. and Parmar, M. K. B. (1993) Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet*, **341**, 418–422.
Tierney, L. (1990) *Lisp-Stat: an Object-orientated Environment for Statistical Computing and Dynamic Graphics*. New York: Wiley.
Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Ass.*, **81**, 82–86.
Young, K. D. S. and Pettit, L. I. (1993) Measuring discordancy between prior and data. To be published.